Effective Hyperspectral Band Selection and Multispectral Sensing Based Data Reduction and Applications in Food Analysis

by

Julius Tschannerl

Centre for Signal and Image Processing

Department of Electronic and Electrical Engineering

University of Strathclyde, Glasgow

A thesis submitted in the fulfillment for the degree of

Doctor of Philosophy

June 5, 2019

Declaration of Authorship

This thesis is the result of the author's original research. It has been composed by the author and has not been previously submitted for examination which has led to the award of a degree.

The copyright of this thesis belongs to the author under the terms of the United Kingdom Copyright Acts as qualified by University of Strathclyde Regulation 3.50. Due acknowledgement must always be made of the use of any material contained in, or derived from, this thesis.

Julius Tschannerl

June 5, 2019

Acknowledgements

Throughout the three years of my PhD journey, there were numerous people who helped me on technical and personal levels to achieve this big undertaking. I want to take the opportunity to thank some of these people explicitly.

Abstract

This thesis focuses on the development of new band selection algorithms and sensing technology to accommodate for the growing demand in commercial and industrial hyperspectral imaging applications. Hyperspectral imaging is inherently complex as it combines two-dimensional spatial image information with highly resolved spectral measurements in a three-dimensional data structure. With rising dimensionality of the data, the required complexity of mathematical models increases exponentially. Dimensionality reduction in the form of band selection is analysed in this thesis, as it preserves physical interpretability of the original data. Two separate approaches are developed. The first one defines an optimisation heuristic for a criterion for band subsets carrying most information and least redundancies. The second approach embeds the identification of the most relevant bands into the reconstruction of the whole dataset. Both approaches successfully reduce the data amount while maintaining the classification accuracy.

This thesis further develops a prototype for a cost-effective, portable and easy-to-use hyperspectral imaging system. RGB LED based time-multiplexed illumination enables rapid acquisition of multi-channel images that are subsequently used to reconstruct spectral signatures from hyperspectral prior via a multilayer perceptron. It is shown that a very accurate reconstruction is possible for a limited number of desired signatures that are sufficiently distinct.

Lastly, the thesis focuses on industrial case studies of hyperspectral imaging in the context of food quality monitoring. The capabilities of SWIR and UV imaging were compared to estimate the smokiness of Scotch Whisky from the concentration of pheno-

lic flavour compounds on barley malt used for distillation. UV imaging shows potential to estimate the concentration but is in its current state of technological development not ready for implementation, as opposed to SWIR, which due to its maturity, has the potential to be implemented in a production environment. Problems arising when imaging food products are examined on the example of salmon fillets.

Contents

List of Figures

List of Tables

Nomenclature

General mathematical notation

x	scalar
$\mathbf{x}$	vector
$\mathbf{A}$	matrix
A	constant
$|x|$	absolute value of scalar x
$|\mathbf{x}|$	norm of vector $\mathbf{x}$
$\langle \mathbf{x}, \mathbf{y} \rangle$, $\mathbf{x}^T \mathbf{y}$	inner product of vectors $\mathbf{x}$ and $\mathbf{y}$
$f(x)$	value of the function f at x
$\hat{f}(x)$	estimate of the value of the function f at x
$\mathrm{e}^x, \exp(x)$	exponential function of x
$\log(x)$	logarithm of x to base 10
$\varphi(\mathbf{x})$	non-linear function mapping $\mathbf{x}$ onto a higher dimensional embedding space
$K(\mathbf{x}, \mathbf{y})$	kernel function realising the inner product of $\mathbf{x}$ and $\mathbf{y}$ in a higher dimensional embedding space

Probability theory

$X, Y, ...$	random variables
$x_1, x_2, ..., x_n$	realisations of random variable X
$p(x)$	probability density function of the random variable X
$p(x, y)$	joint probability density function of the random variable X and Y
$\int_X p(x)dx$	integral of the probability density function $p(x)$ of random variable X
$H(X)$	Shannon entropy of random variable X
$H(X, Y)$	joint Shannon entropy of random variables X and Y
$I(X; Y)$	mutual information of random variables X and Y
$D_{KL}(X \parallel Y)$	Kullback-Leibler divergence between random variables X and Y
$K(\mathbf{u})$	statistical kernel function representing a weighting function for non-parametric density estimation

Graph theory

V	set of vertices
E	set of edges
$G = (V, E)$	graph G consisting of vertices V and edges E

Computational complexity theory

$\mathcal{O}(g(n))$	order of function $f(n) = \mathcal{O}(g(n))$ as $n \to \infty$

Set theory

$A, B, ...$	sets $A, B, ...$
$a, b, ...$	elements of sets $A, B, ...$

Nomenclature

$\forall$	for all
$\exists$	there exists
$\in$	element of
$\bigcup$	union
$\bigcap$	intersection

Optimisation theory

$\max f(x)$	maximum value of function $f(x)$
$\min f(x)$	minimum value of function $f(x)$
$\arg\max_x f(x)$	the value of x that maximises the function $f(x)$
$\arg\min_x f(x)$	the value of x that minimises the function $f(x)$
$\Delta f(x)$	gradient of function $f(x)$
$\frac{\partial f}{\partial x}$	partial derivative of function $f(x, y, ...)$ with respect to x
$\|\mathbf{X}\|_1$	L^1 norm of matrix $\mathbf{X}$
$\|\mathbf{X}\|_2$	L^2 norm of matrix $\mathbf{X}$, also referred to as Frobenius norm
$\|\mathbf{X}\|_{2,1}$	$L^{2,1}$ norm of matrix $\mathbf{X}$
$\boldsymbol{\theta}$	parameter vector
$\mathcal{J}(\boldsymbol{\theta})$	objective function with parameter vector $\boldsymbol{\theta}$
$\Omega(\boldsymbol{\theta})$	regularisation term with respect to $\boldsymbol{\theta}$
η	learning rate

Neural networks

$\phi(z)$	activation function of neuron z

a_i^l activation of neuron i in layer l

$\mathbf{W}^l$ weight matrix of layer l

$\mathbf{b}^l$ bias vector of layer l

$w_{j,i}^l$ weight of neural connection between neoron i in layer $l-1$ and neuron j in layer l

b_i^l additive bias into neuron i in layer l

Hyperspectral imaging specific nomenclature

λ wavelength

$\boldsymbol{I}$ three-dimensional hypercube

$\mathbf{I}_\lambda$ greyscale image at wavelength λ_i of hypercube $\boldsymbol{I}$

$\mathbf{x}_i$ spectrum of i-th pixel in hypercube $\boldsymbol{I}$

$\mathbf{I}_k$ greyscale line image representing k-th line of the hypercube $\boldsymbol{I}$ with all spectral components

Acronyms

Chapter 1

Introduction

1.1 Motivation and aims

Spectral imaging has long been a major research area in the remote sensing community. Sensors such as the Airborne Visible/Infrared Imaging Spectrometer (AVIRIS) [1], Hyperspectral Digital Imagery Collection Experiment (HYDICE) [2] or Reflective Optics System Imaging Spectrometer (ROSIS) [3] collect airborne data in a modality which was then called *imaging spectroscopy*. Typical applications range from land-cover analysis [4] to military target detection [5]. In the past decade, spectral imaging has gained increasing interest outside the remote sensing field, due to the availability of commercial imaging technology and the rapid advances of signal and image processing techniques. The increasing interest in spectral imaging results also in many new requirements such as reduced data amounts and efficient data exploitation algorithms as well as cost and time-effective sensing technology.

Digital imaging traditionally measures the intensity of incident light on a two-dimensional array of sensor elements, resulting in a greyscale representation of the imaged scene. With the use of filters in combination with multiple sensors, or an array of filters on the sensor elements arranged in a so-called *bayer pattern*, colour imaging can be realised. Typically, three channels, *red*, *green* and *blue*, are recorded in alignment with the human visual system [6]. The data can be seen as three images in different colour channels, yielding three intensity values for each pixel. Spectral
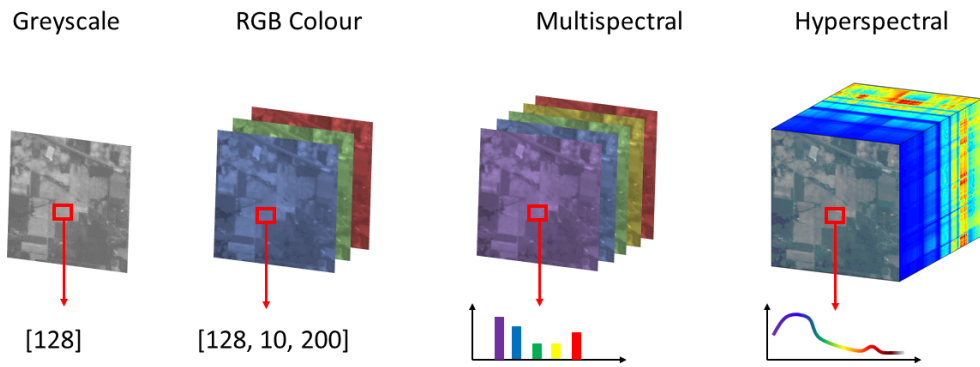
Figure 1.1: From greyscale to hyperspectral imaging.

imaging in contrast aims at increasing the number of recorded channels and extending the range outside the region visible to humans, effectively resulting in spectroscopic measurements at pixel level. Spectral imaging in turn can be divided into Multispectral Imaging (MSI) and Hyperspectral Imaging (HSI), where MSI typically consists of tens of spectral images at distinct wavelengths and HSI delivers a continuous spectrum over a specific spectral region. Figure 1.1 visualises the difference between those four imaging modalities on the Indian Pines dataset (Appendix A.1) recorded by the AVIRIS sensor. Conventional image processing techniques typically operate in a two-dimensional spatial domain. Spectral imaging in contrast combines spectroscopy and digital imaging and as a result operates at the intersection of image and signal processing and spectroscopy. Therefore, many additional considerations and challenges need to be addressed when handling spectral imaging data.

One of the main challenges arising when dealing with spectral imaging data is the amount of data required to store, transmit and process. Where techniques such as image compression [7] can deal with storage and transmission problems, data processing bears additional challenges. Depending on the imaging sensor, spectral data can be

present from tens to hundreds of different channels in various spectral regions such as Ultra Violet (UV), Visible spectrum (VIS) or Short Wave Infrared (SWIR). Depending on the application, usually only a portion of the recorded bands are necessary and redundant information increases the complexity. Techniques such as feature extraction, that operate on the spectral domain and project the feature space onto a lower dimensional subspace have proven very efficient for spectral imaging data [8–10]. Advanced technique even make use of spectral and spatial information for this purpose [11]. The drawback of either approach however is that the projections mask the spectral information of the original data and deprive it off its physical interpretability. Feature selection or band selection in contrast offers a possibility to select only a subset of the spectral bands present in a hypercube, without altering the integrity of the data. Feature selection in the context of spectral imaging has been a recent focus of research with significant results [12, 13]. An inverse approach to data reduction of a high dimensional spectral image is the dimensionality reduction at recording time. Instead of recording high dimensional data and then discarding large amounts of irrelevant data, prior information can be used to only record wavelengths of interest to simplify the recording process. This can lead to faster and more efficient data acquisition as well as decreased cost and complexity of operation.

With the increasing variety of applications of spectral imaging, the demand for more flexible and cost-effective imaging technology is rising. One of the most prevalent imaging techniques for spectral data is line scanning, where either the camera or the image object requires linear translation along the scanning axis. This works well in conjunction with airborne devices or industrial mechanics such as conveyor belts, but limits the application fields dramatically. Another popular scanning technique involves the usage of tunable filters or even the usage of multiple sensor elements, which quickly runs up the cost. Medical applications in particular require systems that are very flexible in their handling and minimally intrusive. Recent developments focus on areas such as snapshot [14, 15] and handheld imaging [16], where the main objective is portability, cost-effectiveness and ease of use.

In addition to the vast amount of remote sensing applications, spectral imaging has

proven to be an extremely valuable tool for food safety and quality assessment. Assuring the quality and thereby safety of food products is quintessential for the industry to increase customer satisfaction and therefore profits but also to avoid the distribution of foodborne pathogens from an environmental and health perspective. The quality of food products is often determined by measures such as the moisture content or protein and lipid levels. Determining the levels of these parameters typically requires destructive chemical techniques that take several days of analysis and are prone to human error in the process. HSI has the advantage of rapid data acquisition and relatively fast data analysis so that it can potentially be used as a nearly real-time, non-intrusive chemometric analysis tool competing with traditional techniques. Likewise, the detection and discrimination of pathogenic bacteria has successfully been implemented in the past [17, 18]. A major drawback of HSI however is the relatively low spatial resolution and optical limitations, leading to often impure measurements which hamper with the accuracy of estimation techniques. The careful planning of controlled experiments as well as the proper choice of HSI systems is crucial to the success.

The work presented in this thesis will address the issues of band selection, efficient data acquisition and analysis by developing algorithms and application driven novel imaging and data analysis approaches. Specifically, the following objectives are defined.

1. Determining optimal models and heuristics that can be used to automatically detect information content of spectral bands in HSI data to subsequently select the most informative band subset.

2. Developing a prototypical imaging system that realises cost-effective, portable and flexible HSI as well as effective models for hyperspectral data reconstruction from multi-channel data.

3. Investigating the potential of HSI in food and drink safety and quality monitoring in two applications: Estimating the smokiness of Scotch Whisky using appropriate classification and regression techniques and detecting listeria bacteria on salmon fillets.

1.2 Original contributions

In this thesis, several different attempts were made to address the main objectives highlighted in Section 1.1. The major original contributions of this thesis can be summarised as follow:

1. **VarBWFastMI for the fast calculation of mutual information in hyperspectral imagery**

   Many band selection algorithms utilise Mutual Information (MI) as a measure for common information between spectral bands. For accurate calculations, Kernel Density Estimation (KDE) methods are preferred for the estimation of the Probability Density Functions (PDFs). A comprehensive analysis of the computational implications for information theoretic measures is provided. To account for the immense computational burden posed by this, Variable Bandwidth Fast Pairwise Mutual Information (VarBWFastMI) is developed as an extension of [19] to incorporate variable kernel bandwidths required for HSI bands. These findings were summarised in a journal article detailed in Appendix B.1.2.

2. **MIMR-DGSA for robust and fast hyperspectral band selection**

   In hyperspectal band subset selection, a typical approach is to combine Evolutionary Algorithms (EAs) for subset generation with an evaluation criterion. In this thesis, a modified Discrete Gravitational Search Algorithm (DGSA) is developed, based on [20], by defining a neighbourhood and movement concept for band subsets and to address the issue of too many hyperparameters present in most EAs. This algorithm is then used to optimise the Maximum Information Minimum Redundancy (MIMR) criterion developed in [13]. The proposed MIMR-DGSA is shown to be faster, more robust and has fewer hyperparameters than other state of the art algorithm. The results were equally presented in the article detailed in Appendix B.1.2.

3. **Segmented sparse autoencoders for hyperspectral band selection**

   An alternative way of approaching band selection are embedded methods that

incorporate the identification of most relevant bands into the training procedure of a machine learning algorithm. Here, this is realised using an Autoencoder (AE) that tries to reproduce spectral signals of an HSI dataset. By posing a sparsity constraint on the input weights of the AE, most relevant bands for reconstruction can be identified. Additionally, the input data is segmented into logical spectral regions to accelerate convergence and achieve a better coverage of the whole spectral range which ultimately leads to a fast algorithm outperforming several state-of-the-art algorithms. The algorithm was presented at EUVIP 2018, as detailed in Appendix B.2.1.

4. **Development of a low-cost, portable hyperspectral camera system using spectral reconstruction from multi-channel images**

   Light Emitting Diode (LED) technology has the advantage of being very small, producing little heat, being capable of generating different wavelengths, and being able to be switched on and off rapidly. This allows it to be used as a time-multiplexed illumination source for a camera with high framerate. Using Red Green and Blue (RGB) LEDs, a multi-channel imaging system is developed that can record up to eleven colour channels at a minimum rate of 18 images per second and a maximum of 200. Based on the multi-channel data, a reconstruction algorithm using Multilayer Perceptrons (MLPs) is developed. The algorithm is able to reconstruct hyperspectral signatures from a minimum of six different colour channels, given sufficient HSI prior, limited required spectra, and sufficiently high discriminability between them. The camera system was originally presented at HSI 2018 and then extended to a journal article as detailed in Appendices B.2.2 and B.1.1.

5. **HSI for the estimation of the smokiness of Scotch Whisky**

   HSI, as a chemometric analysis tool in food quality monitoring is gaining increasing interest. A series of experiments with the aim of determining the concentration of phenolic compounds adhering to the surface of barley grains used for the production of Scotch Whisky is developed. It could be shown that the

mean concentration of a smaller batch of grains can be predicted very accurately
with HSI, given that sufficient control over the lighting and imaging process is
available. This shows the potential of HSI as a real-time, non-destructive tool
competing traditional chemical analysis techniques. These findings were initially
presented at OCM 2017 and the extended experiments and results were published
as a journal article as detailed in Appendices B.2.3 and B.1.4.

1.3   Thesis organisation

The remainder of the thesis is organised as follows:

Chapter 2 provides detailed information on HSI, how the data is constituted, what
differentiates it from MSI data and several techniques of data acquisition in HSI. A lit-
erature review is also conducted focusing on recent developments in the field of hyper-
spectral band selection introducing various techniques for supervised an unsupervised
algorithms. Additionally, relevant recent literature regarding hyperspectral reconstruc-
tion from RGB and MSI data is reviewed and finally, an overview over a variety of
applications of HSI is given, leading to the recent focus on laboratory and industry
based applications.

Chapter 3 introduces the main theoretical concepts and mathematical backgrounds
that lie the foundation of the work in this thesis. Starting with details about Support
Vector Machines (SVMs) for classification and regression, leading to performance and
spectral distance measures, some techniques that are utilised throughout the thesis are
introduced. Then, theoretical basics of band selection are established along with details
about information theoretic measures and general specifications of EAs. Finally, the
mathematical basics of multilayer Neural Networks (NNs) are introduced leading to
the special case of AEs.

Chapter 4 proposes two different band selection algorithms. At first, MIMR-DGSA
is developed building on an information theoretic criterion to evaluate band subsets.
VarBWFastMI is developed to account for varying sizes of kernel bandwidths for the es-
timation of PDFs of spectral bands and to accelerate the calculation time. Afterwards,

DGSA as a discrete EA solving the problem of band subset generation is developed along with detailed analysis of the performance and comparison to various other algorithms.

Using LED technology, Chapter 5 will provide details about a prototyped multi colour channel imaging system based on time-multiplexed illumination. Based on this camera system, a NN approach for hyperspectal reconstruction using HSI prior is proposed that is shown in various experiments to be able to accurately reconstruct spectral signatures.

Chapter 6 finally focuses on applications of HSI in the field of food quality and safety assurance. In this context, the wide range of recent applications of HSI are established laying the foundations for two studies conducted in this thesis. Firstly, an extensive study on barley malt is conducted that tries to estimate the smokiness of the resulting Scotch Whisky from the malt. Secondly, the potential of HSI to detect pathogenic bacteria on the surface of salmon fillets, which proves to pose many practical hurdles, is looked at. Preliminary results will highlight the potential and drawbacks of HSI in this field.

Chapter 7 finally provides some concluding remarks about the work conducted in the context of this thesis and also details plans and possibilities for future work that will extend and improve presented results.

Chapter 2

Background and Related Work

2.1 Introduction

As previously discussed, HSI is distinctly different from RGB imaging in the respect that it aims at delivering spectroscopic measurements on a spatial map as opposed to recreating a scene as the human visual system would perceive it. To establish fundamental terminology and concepts about HSI, firstly, the theory of HSI is established focusing on the differences to MSI. Additionally to data representation, data acquisition techniques are introduced to form a basic understanding of the developed HSI prototype in this thesis. Radiometric calibration and pre-processing techniques of HSI data are introduced here to highlight implications of the specific type of data.

The main focus of this thesis is in hyperspectral band selection, where a comprehensive review of the research in this area is presented, highlighting advantages and shortcomings that are addressed in the developed algorithms in this thesis. Moving on from this, recent advances in hyperspectral reconstruction, with the main goal of developing cost-effective imaging systems are presented, that relate to the developed camera system and reconstruction algorithm developed in this thesis.

Finally, an overview of the vast fields of application of HSI is given which offers the context for the industrial case studies conducted in this thesis.
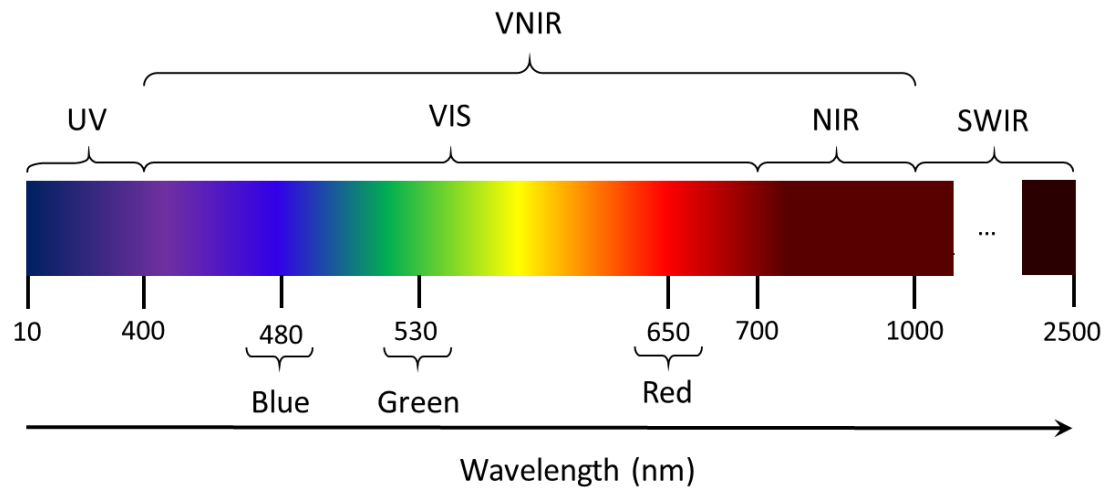
Figure 2.1: Visualisation of spectral regions over a part of the electromagentic spectrum. Key wavelengths roughly dividing the spectral regions defined are indicated. The wavelengths for red, green, and blue are commonly defined as regions and the peak wavelengths are provided for orientation.

2.2   Hyperspectral imaging

2.2.1   Multispectral and hyperspectral data

Computer vision can operate on many different modalities, depending on the specific needs of an application. Classic applications work with monochromatic or RGB colour imaging aimed at reproducing objects on an image plane the way they are perceived by the human eye. Computer vision then applies various processing and analysis techniques to acquire information about the imaged objects. Computer vision aims not only at automating tasks that can be done by humans, but tries to utilise technology to visualise things the human eye cannot perceive. MSI and HSI can be seen as an augmentation of RGB data. The objective is no longer to simply emulate human vision, but to obtain spectroscopic measurements at every pixel of the image, covering different spectral regions within and outside the VIS. HSI is able to capture specific chemical and physical attributes of materials the human eye is unable to perceive. This is a major advantage over common computer vision and enables HSI as a chemometric analysis tool combining most advantages of digital imaging, such as spatial mapping, rapid data acquisition and visualisation with spectroscopic measurements.
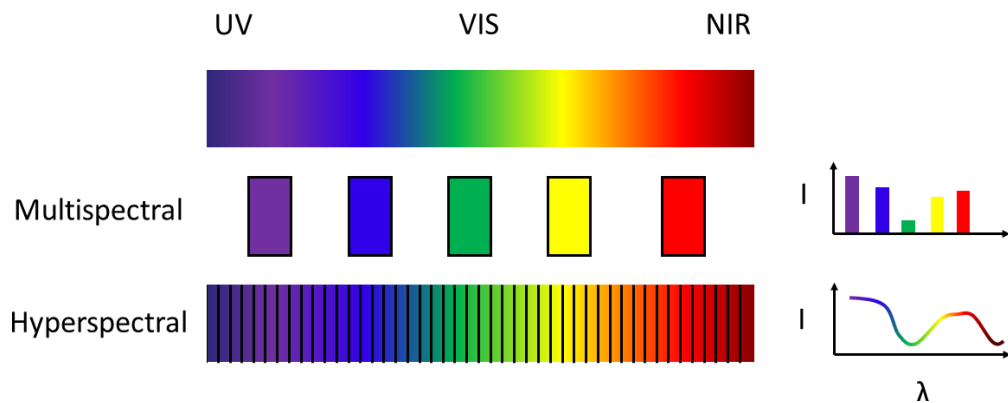
Figure 2.2: Difference between multispectral and hyperspectral imaging. Multispectral data consists of several discrete wavelengths, whereas hyperspectral measures a continuous spectrum.

Common MSI and HSI cameras operate in the regions of the electromagnetic spectrum visualised in Figure 2.1. The UV region typically ranges roughly from 10 - 400 nm, the VIS, where the colours we can perceive as humans lie, ranges from about 400 - 700 nm, the Near Infrared (NIR) is adjacent to that up to 1000 nm and above lies the SWIR up to a maximum of 2500 nm. Some HSI cameras cover the VIS and NIR region, which is typically referred to as Visible to Near Infrared (VNIR). These definitions vary in literature, but for the sake of consistency, the spectral regions are defined as such in this thesis and the ranges covered by the cameras used in this thesis are named explicitly when describing the experiments. What differentiates MSI and HSI are mainly two factors. MSI captures a limited number of discrete spectral wavelengths, typically not more than 20. HSI in contrast incorporates several hundred wavelengths in a continuous spectral range. As a result, MSI data consists of several, distinct spectral measurements, whereas HSI data delivers a continuous spectrum over a certain range, as visualised in Figure 2.2. From a technical perspective, RGB images can be seen as a special case of MSI data. Because of their distinct purpose of depicting colour as perceived by the human eye, they are treated as a separate entity in this thesis. The term *multi-channel* is used in this thesis to describe images with several colour channels based on RGB colour mixing, as opposed to MSI, where distinct peak wavelengths are achieved.

Hyperspectral data can be interpreted as a three-dimensional datastructure, often referred to as a *hypercube*, with two spatial dimensions x and y and one spectral dimension λ. Figure 2.3 visualises such a hypercube, denoted as $\boldsymbol{I}$. $\boldsymbol{I}$ can be viewed at from several view points. $\mathbf{I}_\lambda$ is a two dimensional greyscale image at specified wavelength λ that encapsulates the spatial mapping of the datacube. $\mathbf{I}_k$ in contrast denotes a greyscale image that represents only one spatial line of the image with all spatial components of every pixel of that line. $\mathbf{x}_i$ denotes a single spatial pixel with all its spectral components, i.e. one spectroscopic measurement at pixel level.
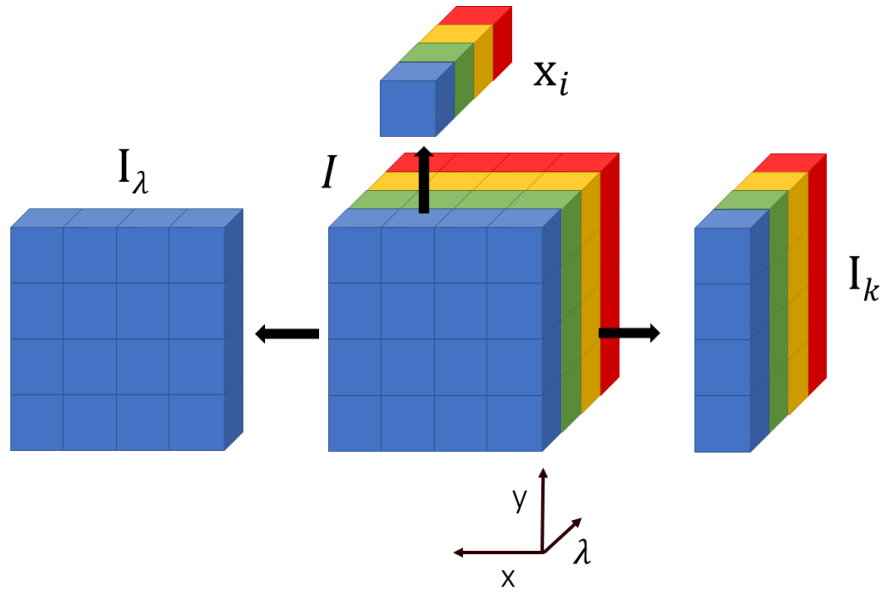
Figure 2.3: Representation of a hyperspectral datacube with all its components.

2.2.2   Data acquisition

There are several ways of acquiring multi- or hyperspectral data, most of which involve a scanning process. Figure 2.4 illustrates three main concepts. Point scanning, also called *whiskbroom scanning*, refers to a mechanism where a single pixel is recorded at a time and an optical dispersive element called *spectrograph* diffracts the incident light into its spectral components which are then integrated on a sensor element consisting of one line of pixels. Line scanning, sometimes referred to as *pushbroom scanning*, denotes a similar technique, but instead of one pixel, a whole line of the image is guided

through an aperture which is then diffracted, resulting in a two-dimensional image representing a single spatial line with every pixel's spectrum and can be integrated by common two-dimensional image sensors. Both these techniques require mechanical steps, where either the camera or the imaged object needs to move along the scanning axis. This requires careful operation to preserve geometric coherence. Pushbroom scanning has proven very compatible with airborne imaging and industrial mechanics, such as conveyor belts and has long been state-of-the-art [21, 22]. Area scanning in contrast sequentially records frames of fully spatially resolved two-dimensional images at different wavelengths at a time. This is typically achieved by using different bandpass filters that require either mechanical devices such as filter wheels or the utilisation of tunable filters such as Acousto-optic Tunable Filters. Geometric distortions are a minor concern, as opposed to effects such as spectral smear. This occurs when the object moves during the scanning process and its position within the image alters between wavelengths. Figure 2.5 compares the principles of line and area scanning visually.
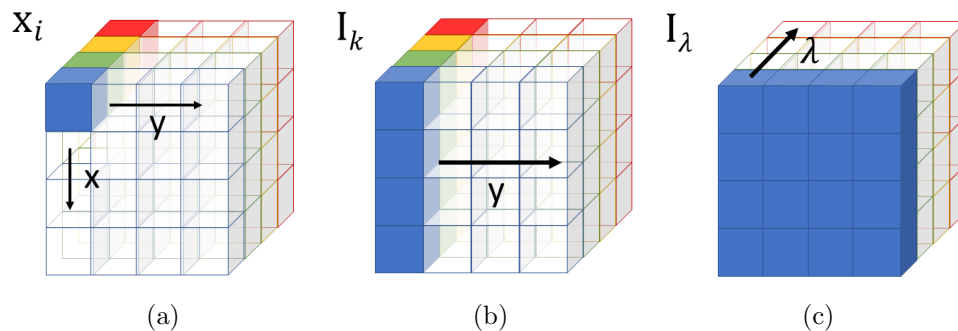
Figure 2.4: Different scanning methods of acquiring hyperspectral data. (a) Point scanning, (b) line scanning, and (c) area scanning. The scanning directions are indicated by the arrows.

For multispectral images with a very limited amount of wavelengths (typically less than ten), it is also possible to utilise several sensor elements in parallel, allowing for the integration in one snapshot, at the cost of additional hardware. Recent research also focuses on the development of MSI and HSI snapshot imagers with technologies such as compressive sensing [23–25], spectral basis multiplexing [15] or novel technologies such as image mapping spectrometers [14]. Coded aperture [26] even allows for Raman

(a)

(b)
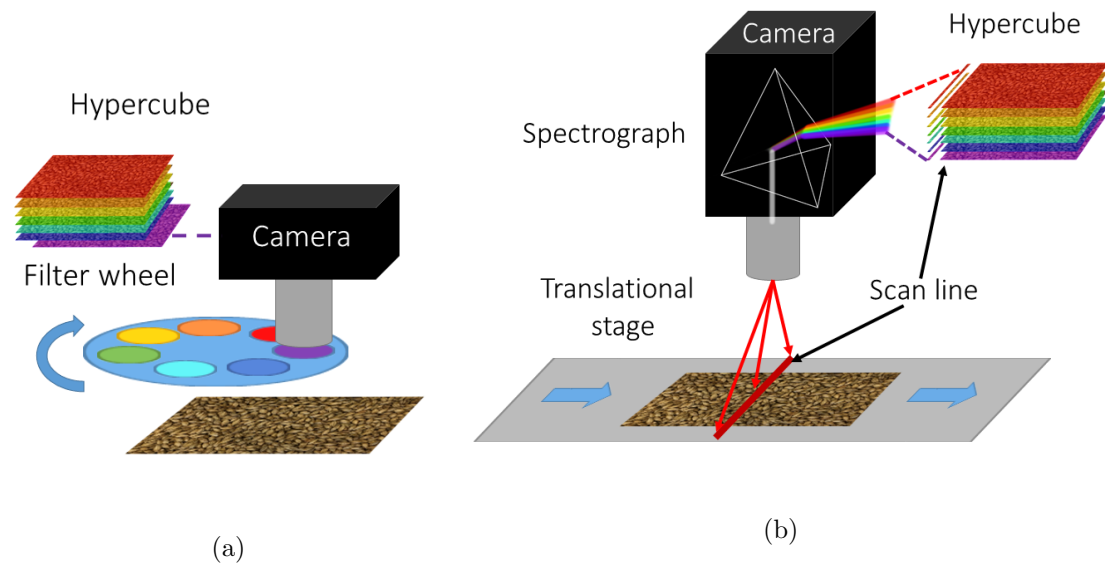
Figure 2.5: Comparison between different scanning techniques. (a) Area scanning and (b) pushbroom scanning. In area scanning, objects need to be as static as possible to avoid spectral distortions, whereas in line scanning, either the object, or the camera needs to move along the scanning axis in a controlled way, to mitigate geometric distortions.

hyperspectral snapshot imaging [27]. Snapshot imaging, however, also suffers frequently from the requirement of very costly and high precision optical components, as well as reduced spectral and spatial resolution and overwhelming amounts of data in a short timespan [28]. The choice of imaging technology is largely dependent on the specific application.

2.2.3   Calibration and pre-processing

In aerial hyperspectral imaging, there are several effects that contribute to the measured spectra. The reflectance attributes of the ground scene as well as the atmospheric attenuation and the radiance of the solar illumination at the given time and location of the images all interfere with each other. Various techniques of atmospheric corrections were developed in the past and are discussed in [29]. Details will not be discussed in

this thesis as the focus lies on laboratory acquired data.

When imaging HSI data in laboratory environments, atmospheric effects do not hamper the data as much as the spectral characteristics of the illumination. Lighting conditions may vary between imaging session and even between individual samples. A common way of accounting for this effect is to convert the measured raw radiance spectra $\mathbf{s}$ to percent reflectance spectra $\mathbf{r}$ by the following formula [30]:

$$\mathbf{r} = \frac{\mathbf{s} - \mathbf{d}}{\mathbf{w} - \mathbf{d}} \times 100\% \tag{2.1}$$

where $\mathbf{d}$ and $\mathbf{w}$ represent the dark and white reference respectively. The dark reference is acquired by imaging without any light exposure to the sensor, which is to estimate the sensor's shot noise. The white reference is acquired by imaging an optimally reflective white surface, which has *lambertian scattering*. i.e. it reflects incident light equally diffuse in all directions over the desired spectral range. The white image can estimate the sensor's light sensitivity to the current illumination and normalises the signal based on that. Using one white and one dark spectrum as a reference does however not account for spatial non-uniformity of the illumination along the scan line in pushbroom scanning or over the two-dimensional image plane in area scanning. As established in [31], it is advisory to use a whole line of white reference to normalise the non-uniformities. Applied to area scanning, a whole three-dimensional reference image is required.

The spectra measured are not only dependent on the chemical absorption but also on physical light scattering due to the surface structure of the objects. Objects often have a very uneven surface and different portions are differently exposed to light. This results in shadow effects as well as varying light scattering attributes. These spectral variations typically manifest themselves in additive or multiplicative components on the base spectra. Various spectral pre-processing techniques are reviewed in [32] to address these problems. For additive effects, it is common to employ spectral derivatives, either 1st or 2nd order. These are most commonly realised with Savitzky-Golay smoothing to minimise noise interferences. A widely used technique to compensate for multiplicative scattering distortions is Multiplicative Scatter Correction (MSC) [32], which estimates

scattering coefficients from a supposedly ideal signal whilst minimising the scatter for each individual signal. MSC is however relatively expensive to compute. One of the most commonly used techniques in HSI is the conversion of the spectra to Standard Normal Variate (SNV) [32]. For a given set of n measured reflectance spectra $R = \{\mathbf{r}_1, \mathbf{r}_2, \ldots, \mathbf{r}_n\}$, the SNV for one spectrum $\mathbf{r}_i$ is calculated by:

$$SNV(\mathbf{r}_i) = \frac{(\mathbf{r}_i - \boldsymbol{\mu})}{\boldsymbol{\sigma}}, \quad \boldsymbol{\mu} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{r}_i, \quad \boldsymbol{\sigma} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\mathbf{r}_i - \boldsymbol{\mu})^2} \qquad (2.2)$$

where μ and σ represents the mean and standard deviation of R respectively. This equals a statistical standardisation and has proven to be a very effective pre-processing technique for hyperspectral data [33, 34].

2.3   Hyperspectral band selection

Hyperspectral data is inherently complex because it contains data in both the spatial and spectral domain in a three dimensional data structure. Effects such as the well known *Hughes Phenomenon* [35] highlight the exponentially rising complexity of the inherent data structure with increasing dimensionality. Data reduction is typically achieved using either feature extraction or band selection. Traditional feature extraction techniques including Principal Component Analysis (PCA) and its variations [36–38], Independent Component Analysis (ICA) [39, 40], Fisher's Linear Discriminant Analysis (LDA) [41] and more recent approaches such as Singular Spectrum Analysis [11], sparse representation [42] or the use of stacked autoencoders [8] have proven very effective in dimensionality reduction but do not preserve the original integrity of the data. Band selection in contrast can provide insight into the intrinsic processes that generate the data [43]. It effectively reduces the data while retaining all relevant information and will be looked at in more detail in this thesis.

2.3.1 Supervised band selection

In HSI, adjacent bands are typically highly correlated [13] and can safely be removed without significant information loss. Equally, not all recorded wavelengths are meaningful for the individual application and are therefore not essential for the predictive power of the system. In [44], an overview of common state-of-the-art supervised band selection algorithms is given. These incorporate different measures such as the correlation coefficient, statistical measures like the Chi-Square distribution and most notably the Minimum Redundancy Maximum Relevance (mRMR) criterion that evaluates features by their individual ability to explain class variables while minimising redundancies based on mutual information. In [45], mRMR has been extended and combined with a forward greedy search. As ground truth data is rarely available for hyperspectral remote sensing data, unsupervised techniques provide a more generic approach and are of greater interest for practical applications. Unsupervised band selection can potentially be applied to an arbitrary HSI dataset, without any prior knowledge about the constitution of the data or ground truth.

2.3.2 Unsupervised band selection

Subset selection based approaches

Some band selection algorithms have been reviewed in [46], where the hyperspectral bands are ranked by measures such as the Shannon entropy or spectral derivatives. Other approaches include band clustering using various similarity measures and selecting representatives [47,48]. Popular similarity measures include information theoretical measures [49] or the correlation coefficient [50]. More sophisticated algorithms try to evaluate a complete band subset rather than individually ranking the features. Generating these subsets is however known to be an *NP*-hard problem [51]. Therefore, EAs are often employed to solve such problems heuristically. In feature selection, popular EA techniques include Particle Swarm Optimisation (PSO) [52,53] and Firefly Algorithm (FA) [54,55]. Both algorithms are population based algorithms, where each solution is represented by a particle or firefly respectively. In PSO, particles move within the solu-

tion space based on their own best position and a global optimal position. FA extends this concept and introduces interaction between all solutions to allow better optimisation. These concepts all define solutions as a binary mask determining the presence or absence in the selected feature subset and therefore implicitly solve the question of the optimal number of selected bands. In [56] and [57], optimised versions of PSO and FA for hyperspectal band selection with a fixed number of bands are proposed. This has the advantage of giving the user power over the size of the band subset. The solutions are encoded as indices of the selected bands. [13] use a different approach named Clonal Selection Algorithm (CSA), where solutions are represented by immune system antigens that clone and mutate based on the quality of the solution. CSA is used to optimise the MIMR criterion. Based on entropy and mutual information, the criterion tries to identify subsets with features that individually carry maximum information (entropy) while minimising the redundancy (mutual information) between them. As demonstrated in [13], MIMR-CSA poses a state-of-the-art unsupervised hyperspectral band selection algorithm that outperforms most existing algorithms. A common problem all of the above mentioned EAs face is the number of control parameters [58] based on the objective function and the constitution of the dataset. CSA in particular requires six parameters with control parameters for mutation, cloning and selection that need individual tuning. In addition, CSA has a relatively high number of evaluations because the cloning can lead to a very large amount of potential solutions.

Deep learning based approaches

With recent advances in the field of deep learning, Zhan et. al [59] propose a wrapper method that utilises a Convolutional Neural Network (CNN) to classify HSI data. Band subsets are generated by segmenting the spectral content into several regions and calculating a newly defined measure called the Distance Density (DD) for each of the segments. Based on the DD, a different number of bands is selected from each segment and the final subset is evaluated by the CNN. Even though the CNN is optimally designed so it does not need to be re-trained for every subset, the algorithm still suffers from high computational cost due to repeated evaluations. Additionally,

wrapper methods are largely dependent on the number and quality of ground truth data. In [60], a technique for improving the classification map that exploits the local smoothness characteristic for hyperspectral data is proposed. Embedded band selection algorithms in contrast incorporate the subset selection into the training of the learning algorithm. In [47] the popular embedded feature selection Least Absolute Shrinkage and Selection Operator (LASSO) for hyperspectral data with good results mainly for a higher number of selected bands was adopted. LASSO, however, can only exploit linear relationships between the input features. A recent research focus on AE feature selection in various fields of machine learning showed the potential of this technique for embedded feature selection as AEs are able to handle any sort of input data and have a strong capability of dealing with non-linear relationships. In [61] a feature selection algorithm based on AEs was introduced. By masking input features, i.e. setting their input weight to 0, and subsequently comparing the reconstruction error between each feature being present or not present, the features that generate the largest difference in the error are considered to be most relevant. In [62] the possibility of AEs for feature selection for facial recognition in digital image data was explored. By putting a sparsity constraint on the input weights, it is possible to identify the features that contribute most to the reconstruction.

2.4 Hyperspectral reconstruction

2.4.1 Reconstruction from multispectral data

Conventional HSI systems are very cost intensive and require relatively complicated operation. Simplifying the imaging process has been the focus of a lot of recent research in HSI. One promising approach is the approximation of HSI data from RGB or MSI. LED technology is a very powerful and low-cost technology that is increasingly interesting for the development of multispectral devices [63]. In [64], an MSI system that utilises a set of LEDs with 17 distinct peak wavelengths in the range of 450 - 990 nm to realise time-multiplexed illumination was proposed, that shows good results in various applications. The use of LEDs makes the system very cost effective and flex-

ible, however the selection of peak wavelengths is hard-wired and lacks a generalised solution. In [65] an LED based system with a considerable broader spectral range of 370 - 1630 nm, in which interpolation is used to recover hyperspectral information from 22 wavelengths was proposed. While the system is portable and low-cost, interpolation runs the risk of overlooking specific spectral absorption peaks of materials not covered by the illumination source.

2.4.2 Reconstruction from RGB data

Hyperspectral recovery from mere RGB images was proposed in [66], in which a sparse dictionary is trained on a database of hyperspectral prior. By matching International Commission on Illumination (CIE) colour coordinates, a transformed RGB dictionary is generated to reconstruct the original HSI information from the RGB images. The system is, however, limited to RGB images and requires prior knowledge regarding the spectral sensitivity of the camera system, which is usually not available. Inspired by the concept of spatial super-resolution, a reconstruction algorithm from RGB images based on deep CNNs was proposed in [67]. The reconstruction results were improved in [68] by designing a simplified CNN with fewer layers which is less prone to overfitting and computationally less expensive. Recent developments in deep learning enable CNNs to be equally less prone to label noise [69] and the introduction of new architectures significantly reduces training complexity [70]. Nonetheless, CNNs evolve with an increasing complexity with respect to their design and often require Graphics Processing Unit (GPU) acceleration to accommodate the excessive computation. Designing shallower networks can reduce the amount of GPU memory required, which, nevertheless, still poses a significant limit on the design.

2.5 Applications of hyperspectral imaging

Traditionally, HSI was developed for remote sensing applications [71, 72]. Many of these applications are owed to military funding, where typical tasks include target and anomaly detection. Target detection refers to the detection of a specific, prior known

spectrum within an HSI image, whereas anomaly detection denotes the detection of spectra that are dissimilar to a certain degree from most of the occuring spectra in an image and therefore does not require any knowledge about the target spectrum [41]. Both are of high interest for military applications, where specific objects can be detected from aerial data. Other remote sensing applications include land-cover analysis [4] with implications for agriculture and land mass mapping. This includes mineralogical mapping of surface soils and outcrops [73], the identification of tree species by estimating the chemical composition of the canopy [73] as well as the study of ocean water especially in coastal regions [74, 75]. Where MSI is able to deliver data used in traditional vegetation indices such as Normalised Difference Vegetaion Index (NDVI), HSI is capable of quantifying biophysical properties of land surface and can be used to detect more detailed information such as photosynthetic pigment or plant water content [76]. This is also very important for fields such as precision agriculture, where the determination of nutrient content of the soil, prediction of crop yield, disease content and biomass estimation are of interest [77, 78].

The recent development of HSI cameras that are more lightweight and affordable and the incorporation of different technologies such as Raman spectroscopy led to a rapid development of lab based applications. Its capabilities of measuring chemical compositions can help analyse drugs and detect counterfeits [79, 80]. The main advantage lies in the non-destructive nature of HSI. This is of huge importance when it comes to the analysis and detection of artwork counterfeits. Paintings in the art world can be of immense monetary value and damaging their integrity is to be avoided at any cost. HSI is particularly valuable here as it has proven to be able to distinguish between different colour pigments and can therefore help identify counterfeits [81, 82]. Likewise, HSI was successfully used to measure the concentration of active pharmaceutical ingredients in tablets [83] and the distributional homogeneity [84]. Another very promising field are medical applications. HSI has proven successful in detecting and classifying cancer cells [85] and particularly at detecting margins during surgery [86] which poses a huge potential for surgical optimisation. Other fields include heart and circulatory pathology as well as retinal foot and diabetes [87]. A further growing field is

forensic analysis, where HSI can help enhancing fingerprints [88] and the identification of chemical signatures and bioagent materials [89].

Another area where HSI is of increasing importance is food quality monitoring [22, 90, 91]. This area will be looked at in detail in Chapter 6.

2.6 Summary

In this chapter, the fundamentals of HSI were introduced. This includes a discrimination between RGB, MSI and HSI and a brief summary of common acquisition techniques for HSI data, along with popular calibration and pre-processing techniques for lab based HSI data. This built the basis to investigate literature of the previous years with respect to the main topic of the thesis, hyperspectral band selection as well as recent advances in the field of hyperspectral reconstruction. Finally, a short overview of the most common fields of application of HSI was given. Based on the presented work, Chapters 4, 5 and 6 will develop new approaches in these fields, where Chapter 3 will give details about the necessary mathematical backgrounds of developed techniques.

Chapter 3

Theoretical Background

3.1 Introduction

This chapter will introduce the theoretical and mathematical foundations as well as terminology for the work developed in this thesis. The previous chapter introduced the basics of HSI and discussed work that was previously done in the areas of band selection, reconstruction of hyperspectral data and summarised common application fields of HSI. The theory presented here will now build the basis for the discussed fields.

First of all, classification of hyperspectral data using SVM is introduced. SVMs are employed throughout this thesis for classification of HSI data as a measure of quality for band selection, spectral reconstruction and in industrial applications. Mathematical background is presented in sufficient detail to understand the usage and implications. Similarly, common performance measures for classification and regression are introduced to develop a unified evaluation system throughout this thesis. Measures for spectral similarity are introduced to be able to evaluate spectral reconstruction.

One of the main contributions of this thesis is based on hyperspectral band selection employing information theory and evolutionary algorithms. Therefore, a theoretical framework for band selection, which helps classify the algorithms developed in this thesis and evaluate them in the wider context of feature selection is established. This allows to then define the mathematical basics of information theoretical measures

along with considerations about their computation as well as the definition of the basic mechanisms of EAs.

Many current machine learning algorithms are based on the concepts of NNs and deep learning. The fundamental terminologies as well as mathematical definitions of MLPs are introduced here as well as training algorithms and their implications. Ultimately, the special case of AEs is treated separately, as their functionality is distinctly different from basic NNs

3.2 Data analysis techniques

3.2.1 Support Vector Machines

SVMs were developed in 1995 [92] and are a popular classification algorithm that gained increasing popularity because of their robustness and relatively easy incorporation of non-linear relationships. The basic idea is that any data from two categories can be separated by a hyperplane, given a sufficiently high dimensional representation of the original features. The hyperplane is chosen to maximise the margin between the classes to promote generalisation using only a small percentage of the original datapoints. The basic concept is visualised in Figure 3.1.

Given a set of n datapoints in an N-dimensional space $\mathbf{x}_i \in \mathbb{R}^N, i \in 1, ...n$ and a corresponding label vector $\mathbf{y} \in \mathbb{R}^n, y_i = \pm 1$, the Hyperplane H_p that separates the points in the two classes can be defined as:

$$\mathbf{w}^T\mathbf{x} + b = 0, \forall x \in H_p \tag{3.1}$$

where $\mathbf{w}$ is a weight vector perpendicular to H_p and b a bias. The distance $d(\mathbf{x}_i, H_p)$ between $\mathbf{x}_i$ and H_p can be calculated by:

$$d(\mathbf{x}_i, H_p) = \frac{|\mathbf{w}^T\mathbf{x}_i + b|}{|\mathbf{w}|} \tag{3.2}$$

To maximise the margin, $\mathbf{w}$ and b are constrained by the following inequalities if all n
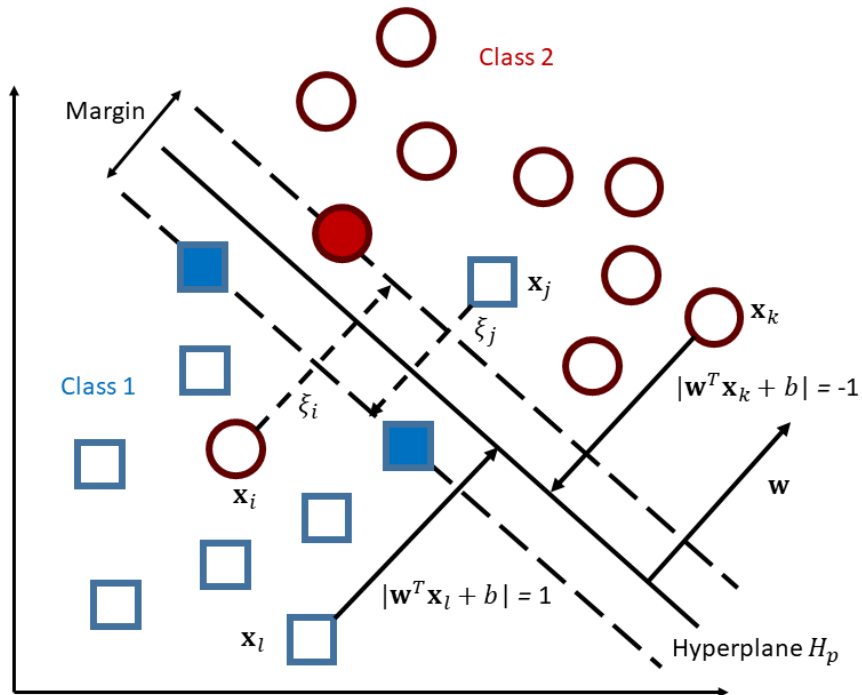
Figure 3.1: Visualisation of the principle of a Support Vector Machine. Datapoints $\mathbf{x}_i$ of two classes are separated by a hyperplane H_p with a margin that maximises the distance between the two classes. The sign of the distance between $\mathbf{x}_i$ and H_p defines the class of $\mathbf{x}_i$. Slack variables ξ_i are introduced to penalise misclassification. One of the main advantages is that only the support vectors (symbolised by filled elements) are required to calculate the classification result.

datapoints are linearly separable:

$$\mathbf{w}^T\mathbf{x}_i + b \geq +1 \quad \text{for} \quad y_i = +1$$
$$\mathbf{w}^T\mathbf{x}_i + b \geq -1 \quad \text{for} \quad y_i = -1, \quad \forall i \in [1, n] \tag{3.3}$$

where the points $\mathbf{x}_i$ that satisfy $\mathbf{w}^T\mathbf{x}_i + b = \pm 1$ lie exactly on the margin and are called the *support vectors*. As shown in Figure 3.1, the margin can be calculated as the distance between two support vectors $\mathbf{x}_i$ and $\mathbf{x}_j$ of different classes $y_i = +1$ and $y_j = -1$ perpendicular to the hyperplane H_p:

$$\frac{(\mathbf{w}^T\mathbf{x}_i + b) - (\mathbf{w}^T\mathbf{x}_j + b)}{|\mathbf{w}|} = \frac{2}{|\mathbf{w}|} \tag{3.4}$$

From this equation it follows that maximising the margin is equal to minimising $\mathbf{w}$. Realising the constraint of Equation 3.3, this can be done for the linearly separable case by the following minimisation:

$$\min \frac{1}{2}|\mathbf{w}|^2, \quad y_i(\mathbf{w}^T\mathbf{x}_i + b) - 1 \geq 0 \tag{3.5}$$

As visualised in Figure 3.1, in reality, most cases are not separable. To account for this fact, *slack variables* ξ_i are introduced. ξ_i penalise misclassification by the following extension of Equation 3.3:

$$\mathbf{w}^T\mathbf{x}_i + b \geq +1 - \xi_i \quad \text{for} \quad y_i = +1$$
$$\mathbf{w}^T\mathbf{x}_i + b \geq -1 + \xi_i \quad \text{for} \quad y_i = -1, \quad \xi_i \geq 0 \quad \forall i \in [1, n] \tag{3.6}$$

Equation 3.5 can then be adjusted to minimise the misclassification by:

$$\min \frac{1}{2}|\mathbf{w}|^2 + C\sum_{i=1}^{n}\xi_i, \quad y_i(\mathbf{w}^T\mathbf{x}_i + b) - 1 + \xi_i \geq 0 \tag{3.7}$$

where C is a weight parameter determining the penalty for misclassification. C is often refereed to as the *complexity* parameter. Via *Lagrange* optimisation, the optimal $\hat{\mathbf{w}}$ and $\hat{b}$ can be determined from Equation 3.7.

The resulting decision function mapping $\mathbf{x}_i$ to the equivalent class y_i is then given by:

$$y_i = \text{sgn}\left(\hat{\mathbf{w}}^T \mathbf{x}_i + \hat{b}\right) \tag{3.8}$$

As previously mentioned, the strength of SVMs lie in their incorporation of non-linearities via mapping functions $\varphi : \mathbb{R}^d \to \mathcal{H}$ that map the datapoints $\mathbf{x}_i \in \mathbb{R}^d$ into the higher dimensional embedding space $\mathcal{H}$. φ must satisfy the *Mercer's condition* [93]. Usually, φ cannot be explicitly defined, but the inner product of two mapped datapoints $\varphi(\mathbf{x}_i)$ and $\varphi(\mathbf{x}_j)$ can be defined by a *kernel function* $K(\mathbf{x}_i, \mathbf{x}_j)$ of the original datapoints $\mathbf{x}_i$ and $\mathbf{x}_j$:

$$K(\mathbf{x}_i, \mathbf{x}_j) = \langle \varphi(\mathbf{x}_i), \varphi(\mathbf{x}_j) \rangle_{\mathcal{H}} \tag{3.9}$$

The kernel function K can be chosen from different types of functions such as linear or polynomial. Here, only the Radial Basis Function (RBF) kernel is of interest, defined as:

$$K(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\frac{|\mathbf{x}_i - \mathbf{x}_j|^2}{2\gamma^2}) \tag{3.10}$$

where γ is an external parameter defining the width of the Gaussian kernel.

From previous definitions, it can be seen that SVM is only defined for a two-class problem. A common method to overcome this issue for k-class problems is the *one-against-all* approach, where $k(k-1)/2$ two-class classifiers are trained with all samples of one class against all samples of all other $k-1$ classes. The final class label of the sample is assigned by the highest number of votes.

SVMs can also be used for regression, usually referred to as Support Vector Regression (SVR), with the slight difference that y_i does take on real values instead of ± 1. The minimisation is then defined as:

$$\min \frac{1}{2}|\mathbf{w}|^2 + C \sum_{i=1}^{n} \xi_i, \quad |y_i - \mathbf{w}^T \mathbf{x}_i - b| \leq \varepsilon + \xi_i \tag{3.11}$$

where ε is a threshold parameter posing an upper boundary for the deviation from the target value. Analogous to Equation 3.8, the regression value y_i of datapoint $\mathbf{x}_i$ can be

obtained by:

$$y_i = \hat{\mathbf{w}}^T \mathbf{x}_i + \hat{b} \tag{3.12}$$

In this thesis, the publicly available library LIBSVM [94] was used for all SVM classification and regression tasks using a Gaussian RBF kernel. Both parameters C and γ were tuned in the range of $\{2^{-10}, 2^{-9}, ..., 0, 2^1, ..., 2^{10}\}$ using 5-fold cross-validation.

SVMs were in the past successfully used for HSI data classification and regression and are still a very popular tool [91, 95, 96]. The popularity of SVMs is among other things due to their very robust mathematical model with a convex optimisation function and easy integration of non-linearities.

3.2.2 Evaluation measures

A common way of measuring the quality of the classification is by examining the *confusion matrix C*. C compares the predicted value of a sample with the actual value of it. An example is given in Table 3.1. On the diagonal, the number of correctly classified samples can be seen. The off-diagonal elements show the number of samples belonging to one class but are classified as a different class. Mathematically speaking, elements $c_{i,i}$ represent the right classification, whereas elements $c_{i,j}$ represent samples known to belong to class i but were classified as class j. Confusion matrices are a useful tool to analyse the performance of a classifier on class level and to identify possible problems.

Table 3.1: Example of a confusion matrix

		Predicted		
		Class 1	**Class 2**	**Class 3**
	Class 1	**15**	1	3
Actual	**Class 2**	0	**30**	8
	Class 3	4	2	**21**

A more general metric to measure the accuracy of a classifier is the Overall Acuracy (OA). The OA sets the number of all correctly classified samples in relation to all

samples. For a dataset with N observations and c classes, the OA is defined as:

$$OA = \frac{1}{N} \sum_{i=1}^{c} n_i \times 100\% \tag{3.13}$$

where n_i represents the number of correctly classified pixels in class i. The OA therefore relates the number of correctly classified pixels to all pixels. This however does not take into account the fact that often the number of samples in each class can vary strongly. The Average Accuracy (AA) in contrast defines the mean of all individual class accuracies calculated by:

$$AA = \frac{1}{c} \sum_{i=1}^{c} \frac{n_i}{N_i} \tag{3.14}$$

where N_i represents the number of samples belonging to class i.

Additionally, Cohen's kappa coefficient can be calculated to assess the classification performance. It quantifies the agreement between the ground truth and the classification results as follows:

$$Kappa = \frac{p_0 - p_e}{1 - p_e} \times 100\% \tag{3.15}$$

where p_0 is the observed level of agreement, identical to the OA, and p_e represents the value expected if the two groups of results, ground truth and classification, were completely independent, defined by:

$$p_e = \frac{1}{N^2} \sum_{i=1}^{c} n_{i_1} n_{i_2} \tag{3.16}$$

where n_{i_1} and n_{i_2} refer to the ground truth and the classification results respectively.

For regression, two popular measures are typically applied to evaluate the quality of the results. For N observations, the Root Mean Squared Error (RMSE) is calculated by:

$$RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (y_i - \hat{y}_i)^2} \tag{3.17}$$

where y_i represents the actual value of the observation and $\hat{y}_i$ the estimated value. It

has the same unit as the estimated value and can be interpreted as the average error made when estimating. As this is scaled to the range of values possible, the coefficient of correlation value r^2 is often employed as an absolute measure for the quality of the model. It is calculated as follows:

$$r^2 = \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N}(y_i - \bar{y}_i)^2} \tag{3.18}$$

where $\bar{y}_i$ represents the mean of all values. r^2 takes on values between 0 and 1, with 1 indicating 100% prediction accuracy of the model.

3.2.3 Spectral distance measures

Several distance measures are defined that compare spectral signals with each other. Three are of interest here. The Euclidean Minimum Distance (EMD) defines the Euclidean distance of two vectors in an N-dimensional space:

$$EMD(\mathbf{x}, \mathbf{y}) = |\mathbf{x} - \mathbf{y}| = \sqrt{\sum_{i=1}^{N}(x_i - y_i)^2} \tag{3.19}$$

The EMD has the main disadvantage of being prone to scaling, i.e. $EMD(a\mathbf{x}, b\mathbf{y}) \neq EMD(\mathbf{x}, \mathbf{y})$. To account for this, the Spectral Angular Mapper (SAM) can be defined as:

$$SAM(\mathbf{x}, \mathbf{y}) = \arccos\left(\frac{\langle \mathbf{x}, \mathbf{y} \rangle}{|\mathbf{x}||\mathbf{y}|}\right), \quad 0 \leq SAM \leq \frac{\pi}{2} \tag{3.20}$$

SAM is equal to the cosine distance of two vectors and has the attribute of being scaling invariant, i.e. $SAM(a\mathbf{x}, b\mathbf{y}) = SAM(\mathbf{x}, \mathbf{y})$, which is very useful when comparing spectra, as additive or multiplicative distortions that typically occur with varying lighting conditions or shadow effects should be neglected when comparing spectral signatures.

Based on the concept of divergence in information theory, Spectral Information Divergence (SID) was proposed in [97]. Each spectrum $\mathbf{x}$ is modelled as a random

variable defined by a PDF. For two spectra $\mathbf{x}$ and $\mathbf{y}$, the SID is defined as:

$$SID(\mathbf{x}, \mathbf{y}) = D_{KL}(\mathbf{x} \parallel \mathbf{y}) + D_{KL}(\mathbf{y} \parallel \mathbf{x}) \tag{3.21}$$

where $D_{KL}(\mathbf{x} \parallel \mathbf{y})$ denotes the *Kullback-Leibler divergence* or *relative entropy* of a discrete variable $\mathbf{x}$ with respect to $\mathbf{y}$ defined by:

$$D_{KL}(\mathbf{x} \parallel \mathbf{y}) = \sum_{i=1}^{N} x_i \log \left(\frac{x_i}{y_i} \right) \tag{3.22}$$

According to [97], SID outperforms both EMD and SAM in terms of discriminability. SID is a similarity measure, where larger values indicate greater similarity, whereas SAM and EMD calculate a geometric distance, which results in smaller values for greater similarity.

3.3 Information theory

The fundamental unit of information in information theory is the Shannon entropy [98]. It measures the amount of uncertainty a random variable carries. The Shannon entropy $H(X)$ of a continuous random variable X is defined as:

$$H(X) = - \int_X p(x) \log p(x) dx \tag{3.23}$$

where $p(x)$ denotes the PDF of X. The higher the entropy, the higher the uncertainty and therefore the information content of a random variable.

The information shared by two random variables X_1 and X_2 can be measured by the MI $I(X_1; X_2)$, which is defined as:

$$I(X_1; X_2) = \int_{X_1} \int_{X_2} p(x_1, x_2) \log \frac{p(x_1, x_2)}{p(x_1)p(x_2)} dx_1 dx_2 \tag{3.24}$$

where $p(x_1, x_2)$ is the joint PDFs of random variables X_1 and X_2. The higher the value of MI, the more information is shared between the random variables. Two completely independent variables have an MI of zero.

MI can alternatively be expressed by the entropy with the following formula:

$$I(X_1; X_2) = H(X_1) + H(X_2) - H(X_1, X_2) \qquad (3.25)$$

where $H(X_1, X_2)$ denotes the joint entropy of X_1 and X_2.

It is worth noting that both measures depend on the PDFs of random variables and not on the specific values they take on. The entropy and MI both are very popular in various computer vision applications such as feature selection [99–103] or image registration [104, 105].

To minimise the computational cost of calculating the PDFs, many algorithms approximate the PDF by the grey value histogram [102, 103]. These approximations however suffer from quantisation errors. To avoid this, a popular method of estimating the PDF of a random variable is KDE, often referred to as *Parzen window* estimation [106], named after Emanuel Parzen. Given a set of n observations $x_1, x_2, ..., x_n$, the underlying distribution with PDF $p(x)$ can be approximated by:

$$\hat{p}(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right) \qquad (3.26)$$

where $K(.)$ denotes the kernel function or Parzen window and is assumed to be a symmetric PDF. h represents the kernel width, or bandwidth, which controls the smoothness of the resulting density estimate. The choice of h is crucial to the quality of the estimate [107]. Bandwidth selection has been of major research interest over the last decades. In [107, 108] for instance, several methods of automatic bandwidth selection are introduced. The optimal bandwidth ideally is wide enough to smooth out small variations in the data but detailed enough not to over-smooth certain features of the distribution. Figure 3.2 illustrates the effects of different bandwidths in a density estimation for a Gaussian mixture. While wider bandwidths are able to reproduce the smoothness of the original function, they fail in capturing details, particularly visible at the spike around 0. Lower handwidths are able to recreate the result better but tend to over-adapt to the samples in regions that expose a wider, smooth distribution,

e.g. between 0.5 and 4. Therefore, adaptive bandwidth estimation methods have been developed to capture local variations in the shape of the PDFs [109, 110].
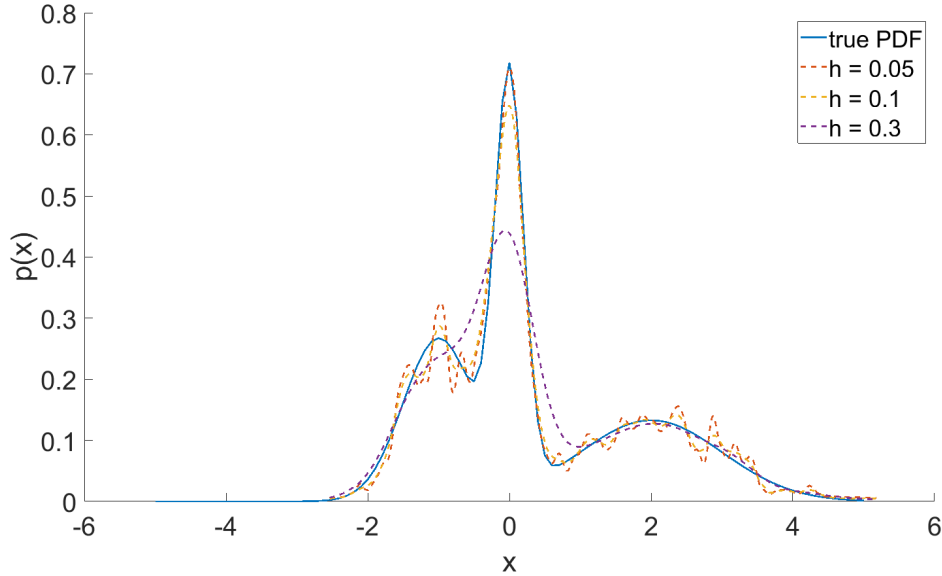
Figure 3.2: Kernel density estimation with different bandwidths. Smaller bandwidths are able to pick up more detail but tend to over-adapt to small changes. Wider bandwidths are able to properly smooth wider regions but tend to over-smooth particular details.

Other than the bandwidth, the choice of the kernel function also has a significant effect on the quality of the estimation. Some of the more popular kernel functions are displayed in Table 3.2. The focus here lies on the Gaussian kernel as this the most commonly used. Combined with Equation 3.26, the PDF of a univariate random variable with Gaussian kernel can be estimated from n datapoints with given bandwidth h by:
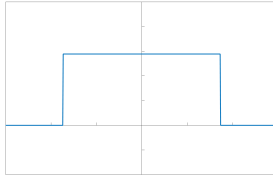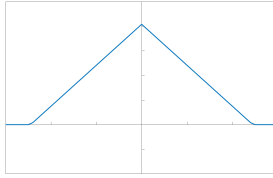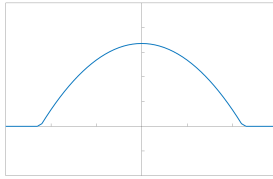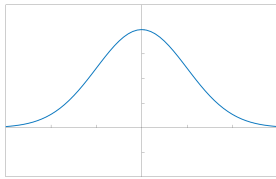
$$\hat{p}(x) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{\sqrt{2\pi h^2}} \exp\left(-\frac{(x-x_i)^2}{2h^2}\right) \tag{3.27}$$

And for a bivariate distribution:

$$\hat{p}(x,y) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{\sqrt{2\pi h^2}} \exp\left(-\frac{(x-x_i)^2 + (y-y_i)^2}{2h^2}\right) \tag{3.28}$$

With these estimates, both the entropy and MI can be approximated using Equations

Table 3.2: Examples of five commonly used kernels alongside a visualisation of their plot

Rectangular $\qquad$ $K(u) = \frac{1}{2}$ for $|u| \leq 1$

Triangular $\qquad$ $K(u) = (1 - |u|)$ for $|u| \leq 1$

Epanechnikov $\qquad$ $K(u) = \frac{3}{4}(1 - u^2)$ for $|u| \leq 1$

Gaussian $\qquad$ $K(u) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}u^2\right)$

3.23 and 3.24.

3.4   Feature selection

Some theoretical basics of feature selection algorithms are introduced in this section. This includes a classification of their evaluation methods and an abstract scheme for the procedure of most common algorithms. Feature selection is introduced in general here. In the context of HSI, it equals band selection and will in the rest of the thesis be referred to as such.

3.4.1 Procedure of feature selection algorithms

Very basic feature selection algorithms perform an individual ranking on each feature and then select the ones with the highest scores. More sophisticated selection algorithms also take into account the interactions between the features as several features can e.g. score high for discriminating different classes but carry redundant information between each other. This is why state of the art selection algorithms select and evaluate a whole subset. According to [111], feature selection generally takes place in four steps.

1. **Subset generation**

   An exhaustive search for an optimal feature subset out of set a of features is known to be an *NP*-hard problem [51]. A heuristic subset generation algorithm that explores all potential combinations efficiently is therefore mandatory. Since we are working with iterative algorithms, there are two issues that need to be addressed. *Subset generation* and *search organisation*. Subset generation determines the strategy with which the successor subset(s) for the next iteration are generated, such as *greedy* or other strategies. The search organisation in turn determines which of the generated successors are investigated.

2. **Subset evaluation**

   The generated subset then need to be evaluated for fitness. According to [111], evaluation criteria can be categorised in two groups; *independent* and *dependent criteria*. Independent criteria exploit the intrinsic characteristics of the training data, independent of the learning algorithm applied later. This can happen in an *supervised* or *unsupervised* manner, depending on the availability of ground truth data. Dependent criteria in turn are defined by the predetermined learning algorithm applied on the data. If the data is subject to a classification task, the accuracy of the chosen classification algorithm determines the fitness of the selected subset.

3. **Stopping criteria**

   The point where the algorithm terminates can be determined by various factors

such as a predefined minimum error, a maximum number of iterations or even by performing an exhaustive search over the entire search region.

4. **Result validation**

   The selected feature subset can be validated for a specific purpose by applying a specific machine learning algorithm. For subset evaluation with a dependent criterion, this is done implicitly, but independent criteria can further be validated this way.

3.4.2 Classification of feature selection methods

With respect to the evaluation methods, literature generally discriminates between three types of feature selection algorithms [43, 111].

1. **Wrapper methods** are dependent on a selected learning algorithm. The fitness of the feature subset is determined by the performance of the chosen learning algorithm. Wrapper methods optimise for the specific learning algorithm that is used to evaluate and therefore lack generality with respect to other algorithms. They also tend to be computationally intensive, as the learning algorithm has to be performed for each generated feature subset. Due to the high dimensionality of hyperspectral data, these methods are very impractical.

2. **Filter methods** in comparison define a substitute criterion for evaluating the fitness of the feature subset. As mentioned in Section 3.4.1, many different criteria, e.g. distance, probabilistic or information theoretical measures, have been defined in the past. The advantage of this method is the focus on yielding a feature subset that performs well for many different applications as it is not tailored to a specific learning algorithm. They are also computationally less complex than wrapper methods as only a fitness function needs to be evaluated rather than performing a full training and testing on the dataset.

3. **Embedded methods**, like wrapper methods, are tailored to a specific learning algorithm. The difference however is that embedded methods incorporate the

feature selection in the definition of the learning algorithm. Similar to wrapper methods, embedded methods optimise for the specified learning algorithm and lack generality but are far less computationally complex.

3.5 Evolutionary algorithms

EAs describe a framework for stochastic algorithms inspired by biological evolution that are commonly used to solve computationally expensive optimisation problems heuristically. A common classification of sub-classes of EAs are Genetic Algorithms (GAs), Evolutionary Strategies (ESs) and Evolutionary Programming (EP) [112] where differences mainly occur in technicalities such as data representation. The basic principle of EAs is the same for all algorithms [113]. A set of candidate solutions, often referred to as *agents*, is defined in an initial population. The quality, or *fitness* of the agents is determined by a pre-defined objective function. Based on this fitness, a selection process takes place where only the best solutions survive. These solutions then undergo *mutation* and *cross-over*. Cross-over, sometimes referred to as *recombination*, refers to the communication between agents, i.e. the agents interact and create new agents or alter themselves. Mutation means the solutions randomly change certain elements of their solution to create variability and a stochastic element. An additional selection process selects only the best solutions and the whole process is repeated until some convergence criterion is satisfied. The basic procedure is visualised in Figure 3.3.

EAs can be classified as so called *generate-and-test* algorithms [112] and as such follow two main principles; *exploration* and *exploitation*. According to [112], these terms are not unambiguous in literature, but here, the notion that they are opposing forces that need to be balanced is followed. Exploration describes the principle of optimally distributing candidate solutions over the entire search space on a global level so that most possibilities are covered and the optimal solution is not left out ideally. It is commonly achieved by an effective initialisation strategy and particularly by mutation and cross-over. Exploitation on the contrary is achieved by mining existing information and is commonly achieved by the selection process. Over the runtime of an EA, it is
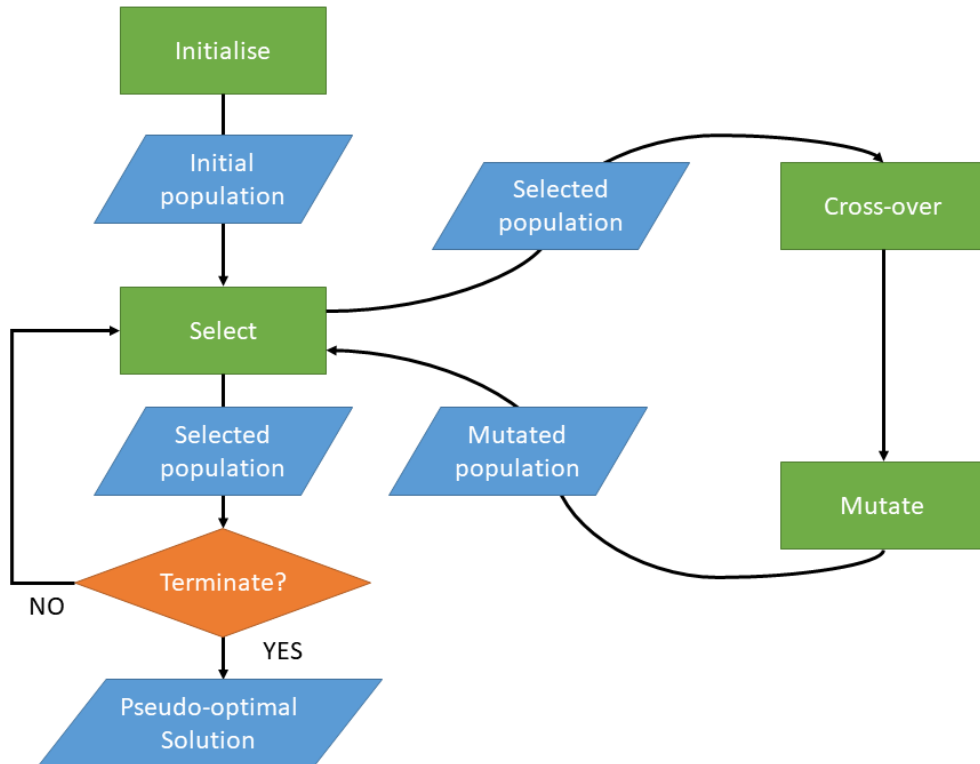
Figure 3.3: General Procedure of evolutionary algorithms. In concrete cases, the order of cross-over and mutation can be altered. The algorithm terminates when a certain criterion is satisfied, e.g. a maximum number of iterations or a minimum error rate.

expected that with convergence, exploration decreases and exploitation increases.

The realisation of each individual step is largely dependent on the nature of the problem to be solved. EAs can be used to solve continuous or discrete problems. Mutation and cross-over do not always have to happen in conjunction. Some algorithms may only define one or the other or alter the order.

3.6 Multilayer neural networks

3.6.1 Fundamentals

A Multilayer neural network, or MLP, consists of a set of connected units, usually referred to as *neurons* because their functionality is loosely based on neurons in the human brain. The neurons are arranged in *layers*, with at least one input and one output layer and an arbitrary number of hidden layers. Neurons are connected via neural connections, or *synapses*. In fully connected layers, each neuron is connected with each neuron of the following layer, where each synapse is multiplied by a weight and all incoming synapses to one neuron are added up. An additional additive bias term is typically introduced before each fully connected layer. The result of all these inputs to a neuron is then subject to an *activation function*, determining the output of the neuron. Figure 3.4 visualises a simple architecture of a basic MLP. Activation functions are mainly introduced to map an input to a neuron, which can theoretically be arbitrary in scale, to a desired range of the specified task. The activation of the output layer determines the result of the NN. This basic architecture allows for NNs to be used in almost any arbitrary machine learning task and provides a relatively simple way of learning non-linearities from input data [114].

In mathematical terms, a NN consists of l layers with neural activations a_i^l, weights $w_{j,i}$, bias terms b_i^l and activation functions ϕ^l. The activation a_j^{l+1} of the j-th neuron in the $(l+1)$-th layer can be calculated by the sum of all d incoming neural connections:

$$a_j^{l+1} = \phi^l \left(\sum_{i=1}^{d} a_i^l w_{j,i}^l + b_j^l \right) \tag{3.29}$$
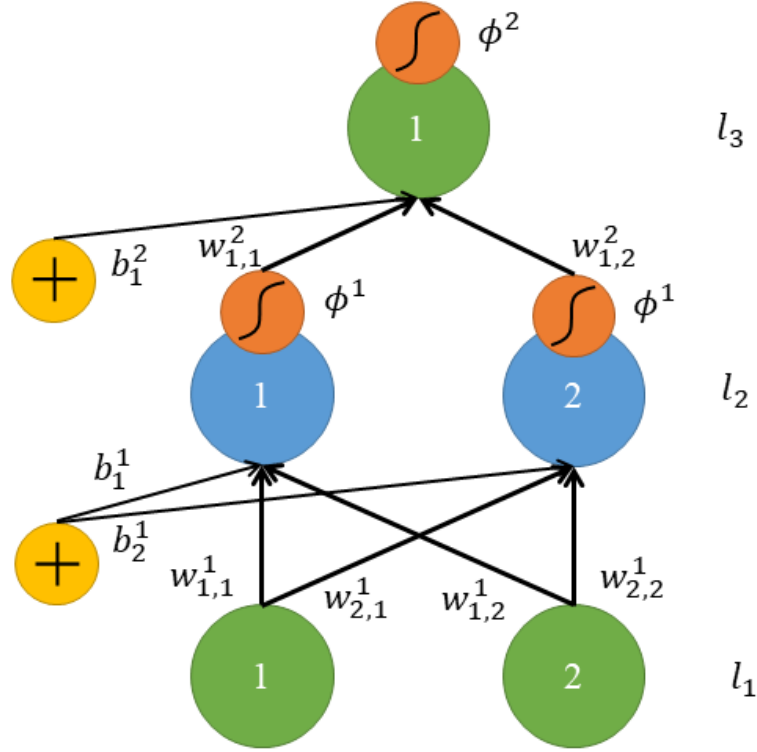
Figure 3.4: Architecture of a basic multilayer neural network. Input and output layers are denoted by l_1 and l_3 respectively and hidden layer by l_2. Neural connections are indicated with the corresponding weights. The bias terms of the according layers are indicated as yellow nodes and the activation functions of the neurons as orange nodes.

where d denotes the number of neurons in layer l, $w_{j,i}^l$ the weights from the i-th neuron in the l-th layer to the j-th neuron in the $(l+1)$-th layer, b_j^l the bias term into the j-th neuron in the $(l+1)$-th layer and ϕ^l the activation function of the $(l+1)$-th layer. This can be rewritten in vectorised form for the whole layer as:

$$\mathbf{a}^{l+1} = \phi^l(\mathbf{W}^l\mathbf{a}^l + \mathbf{b}^l) \tag{3.30}$$

Commonly applied activation functions are listed in Table 3.3. The linear activation function simply performs a linear transformation of the input and if the whole network consists of these, then the network is incapable of non-linear transformation and basically equals a linear regression model [114]. In the simplest form, non-linear transformation can be achieved by thresholding the output of a neuron, i.e. applying

a binary step function. This enables for example binary classification. However, the gradient of this function is zero, which is why it is of no practical use for backpropagation, as detailed in Section 3.6.2. A very popular way of introducing non-linear transformation is the use of sigmoid activation functions. The gradient for this function is the highest for low values, meaning that small changes in this region have strong effects, which make it very useful for classification tasks [115]. The tanh function is in effect a scaled version of the sigmoid function and allows for negative values to be generated. One of the most popular activation functions these days is Rectified Linear Unit (ReLU), which is a linear transformation for positive values and cuts off negative values. This makes the function non-linear and has the advantage of not activating all neurons at one time, resulting in a sparser version of the network making it more efficient to train [116]. These are just a small subset of all available activation functions highlighting general trends.
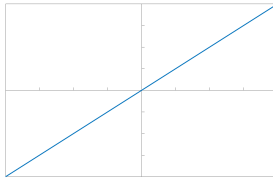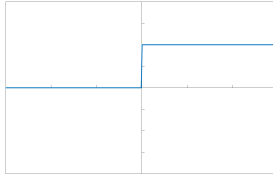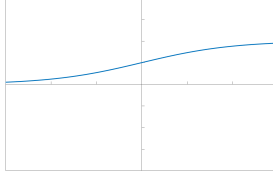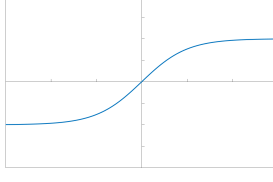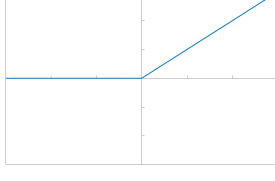
3.6.2   Training of neural networks

A common method to train neural networks is to define an error between the desired and achieved output and alter the weights and biases until this error is minimised. The most popular algorithms applied are variations of the *gradient descent algorithm*. The gradients are calculated using the *backpropagation* algorithm, made famous in [117]. The algorithm will only be described in principle here and details can be looked up in [114].

An untrained network is presented with a set of pairs $(\mathbf{x}, \mathbf{y})$ of input vectors $\mathbf{x}$ and corresponding output vectors $\mathbf{y}$ which then calculates the output $\hat{\mathbf{y}}$. The calculation of an output of the network through all layers of the network is referred to as a *feedforward* operation. After each feedforward operation, the *training error* δ can be calculated with a loss function $\mathcal{J}(\boldsymbol{\theta})$. Given the weights $\mathbf{W}^l$ and biases $\mathbf{b}^l$ with $\boldsymbol{\theta} = \{\mathbf{W}^1, ... \mathbf{W}^n, \mathbf{b}^1, ... \mathbf{b}^n\}$ of a network with n layers, the NN can be defined as a loss function of the parameters $\boldsymbol{\theta}$:

$$\mathcal{J}(\boldsymbol{\theta}) = \frac{1}{2}\|\mathbf{y} - \hat{\mathbf{y}}\|_2 \tag{3.31}$$

Table 3.3: Examples of five commonly used activation functions in NNs alongside a visualisation of their plot

Linear $\phi(z) = z$

Binary step $\phi(z) = \begin{cases} 0 & z < 0 \\ 1 & z \geq 0 \end{cases}$

Sigmoid $\phi(z) = \sigma(z) = \frac{1}{1+e^{-z}}$

Tanh $\phi(z) = \tanh(z) = \frac{(e^z - e^{-z})}{(e^z + e^{-z})}$

ReLU $\phi(z) = \begin{cases} 0 & z < 0 \\ z & z \geq 0 \end{cases}$

The network is trained by adjusting the weights and biases of each layer until the training error converges. These are adjusted by calculating the gradients, or partial derivatives $\Delta\mathbf{W}^l$ and $\Delta\mathbf{b}^l$ at each layer l with:

$$\Delta\mathbf{W}^l = -\eta\frac{\partial\mathcal{J}}{\partial\mathbf{W}^l}$$
$$\Delta\mathbf{b}^l = -\eta\frac{\partial\mathcal{J}}{\partial\mathbf{b}^l}$$

(3.32)

where η denotes the learning rate specified as a hyperparameter of the algorithm. Back-propagation then refers to the calculation of the output error δ and then step by step calculating the error $\boldsymbol{\delta}^l$ at each layer recursively backwards through the network. For the gradient descent, the gradients of the error function with respect to the parameters $\Delta\mathbf{W}^l$ and $\Delta\mathbf{b}^l$ are required. These can be calculated with the corresponding $\boldsymbol{\delta}^l$ and help to adjust the parameters and ultimately enforce the convergence of the network.

A problem that arises when training networks is their generalisation capabilities. The design of networks is in theory arbitrary with respect to the number of hidden layers and neurons, and therefore free parameters. This can cause significant overfitting when choosing a too complex structure or underfitting when the number of neurons cannot sufficiently represent the complexity of the data. Often, simply choosing the right number of layers and neurons is however not sufficient to obtain the desired outcome. This is where regularisation comes into play. A more complex, well regularised network might in fact perform much better than an unregularised optimised network [118]. There are many different ways of applying regularisation to a NN which can be looked up in detail in [118]. Some of the main concepts are explained here.

A very common way to regularise is to introduce parameter norm penalties, with the most popular way being L^2-regularisation. Typically, regularisation is only applied onto the weights of the MLP [118], since the biases require less data to be fitted and regularising both may lead to significant underfitting. In L^2-regularisation, a term $\Omega(\boldsymbol{\theta})$ for all n layers is added to the loss function, defined as:

$$\Omega(\boldsymbol{\theta}) = \frac{\alpha}{2} \sum_{l=1}^{n} \|\mathbf{W}^l\|_2 \tag{3.33}$$

where α is a hyperparameter that controls the effect of the regularisation. L^2-regularisation has the effect of penalising the magnitude of the weights and therefore smoothing the output and enforcing generalisation. This method is often referred to as *weight decay*. Applying the L^1 norm is another, less popular method which has qualitatively a very distinct effect from L^2 regularisation. Applying the L^1 norm results in solutions where the weight parameters are more sparse [118] which has been

extensively used for feature selection algorithms such as LASSO [119].

Other ways of regularisation include augmenting the data, i.e. synthetically generating more training data to improve prediction accuracy. Augmentation can also mean to add noise to the data to increase noise robustness of the model. This may not only include adding noise to the input data, but also to the weights or even the outputs. One specific way of adding noise that will be mentioned here is *dropout*. Dropout is a computationally inexpensive way of approximating *bagging*, which describes the process of combining the results of several separately trained models [118]. It is achieved by effectively generating a binary mask of randomly sampled values multiplied with a non-output layer to remove certain nodes. The probability at which nodes are kept is called the *keep rate* and is another hyperparameter of the network. Dropout roughly doubles the number of iterations required for convergence [120]. A very specific type of NNs are AEs, which are introduced in the following section.

3.7 Autoencoders

A basic AE model is a special feedforward neural network with one input layer and two fully connected layers. Its purpose is to reconstruct the input at the output layer by learning a lower dimensional, abstract representation of the data at the hidden layer. A simple AE based on [62] is visualised in Fig. 3.5. For an input matrix $\mathbf{X} = \{\mathbf{x}_1, ..., \mathbf{x}_m\}^T \in \mathbb{R}^{m \times d}$, where m is the number of input samples and d is the dimensionality of the input, an AE is defined by two functions. The *encoder* function $\mathbf{f}_i = \phi^1(\mathbf{W}^1 \mathbf{x}_i + \mathbf{b}^1)$ and the *decoder* function that reproduces the input $\hat{\mathbf{x}}_i = \phi^2(\mathbf{W}^2 \mathbf{f}_i + \mathbf{b}^2)$. ϕ^1 and ϕ^2 are the activation functions of the hidden and output layer respectively, $\mathbf{W}^1$ and $\mathbf{W}^2$ represent the weight matrices and $\mathbf{b}^1$ and $\mathbf{b}^2$ the bias vectors for hidden and output layer with the same elements as specified in Section 3.6.

For training, the AE can be defined as a loss function $\mathcal{J}(\boldsymbol{\theta})$ of the difference between the input and output with parameter $\boldsymbol{\theta} = \{\mathbf{W}^1, \mathbf{W}^2, \mathbf{b}^1, \mathbf{b}^2\}$.

$$\mathcal{J}(\boldsymbol{\theta}) = \frac{1}{2m} \|\mathbf{X} - \hat{\mathbf{X}}\|_2 \tag{3.34}$$
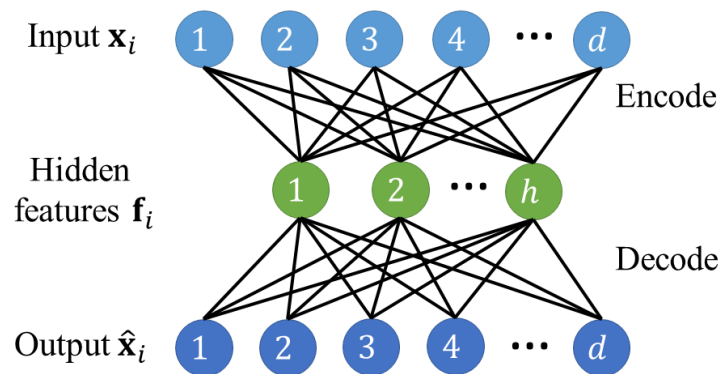
Figure 3.5: Schematic of a basic autoencoder. Input and output have the same dimensionality, whereas the hidden layer can have an arbitrary dimensionality.

Training is commonly done with the backpropagation algorithm specified in Section 3.6.

Autoencoders have several attributes that make them interesting for various applications. The most straight forward application would be data compression, where a lower dimensional representation of the encoded data serves as the basis for a lossy compression of high-dimensional data [121]. Other applications include de-noising [122], generative modelling [123] and notably feature extraction [8] and selection [61,62], which will be examined here in detail.

3.8 Summary

This chapter introduced the mathematical foundations of concepts used in this thesis. At first, the backgrond of SVM classification and regression are shown in detail, as this is the main algorithm used in most applications in this thesis. To evaluate the results of classificatoin and regression, several performance measures were introduced. Additionally, some similarity measures for spectral signals were introduced and briefly discussed with respect to their applicability to HSI data. As a basis for the developed band selection algorithms, the fundamental units of information theory were introduced, including Shannon entropy and MI as well as considerations regarding their calculations. Theoretical basics for evolutionary algorithms were also established that allow to categorise the developed and analysed algorithm in this thesis. Finally, the basics of NNs

in the form of MLPs were introduced including terminology, mathematical definitions and training methods, leading to the specific case of AEs.

Chapter 4

Advanced Techniques for Hyperspectral Band Selection

4.1 Introduction

Hyperspectral band selection, as opposed to common feature extraction techniques such as PCA or ICA that generate new features by linearly combining the original ones, has the advantage of retaining physical interpretability of the underlying data. The problem of high redundancy and vast amount of often superfluous information in HSI datasets for specific applications can efficiently be solved with band selection. Depending on the constitution of the data, band selection might even perform better than feature extraction [124]. While feature extraction performs better for data with high information density between the features [124], it has also been argued that techniques such as PCA are not necessarily suitable for HSI data because linear combinations of features sometimes tend to amplify variations within classes rather than between them [125]. Selecting a subset of the originally available features avoids this problem. Often, contextual knowledge about specific absorption characteristics of imaged object can help to identify spectral regions of interest. When this information is not available, the need for more generic approaches arises.

One of the main challenges in band selection is the determination of information

content of individual spectral bands and their relevance with respect to specific applications. Supervised methods, where label information is available, have the advantage that they can clearly optimise for a specific task as the label information helps identify possible discriminatory features. Unsupervised methods in contrast are considered harder problems because of the lack of this information [126]. On the other hand, this makes unsupervised methods much easier to apply on any arbitrary dataset as label information on datasets is in reality hardly available or only gained with significant effort. Not relying on label information also eradicates the bias of the algorithm towards a specific applications and in turn reveals intrinsic information about the underlying data. In HSI, this effect can be harnessed to gain information about the spectral constitution of the imaged objects and help identify relevant wavelengths.

In this chapter, two approaches to hyperspectral band selection are developed. The focus for both approaches was on unsupervised methods, because they can potentially be applied on arbitrary HSI datasets without any prior knowledge and deliver a reduced dataset containing the most relevant information. The first algorithm analyses the capability of information theory as a measure for the information content of hyperspectral bands. The advantage of these measures is that the results are transparent and the decisions the algorithm makes can be tracked and might reveal information about the underlying data. Based on the MIMR criterion for band subset evaluation developed in [13], the advantages and drawbacks of information theory with respect to their ability to represent information content and practical problems that arise when computing these measures are analysed in detail. Generating band subsets that are evaluated this way is considered an *NP*-hard problem [51], which is why the applicability of various EAs is analysed with respect to their capability of optimising this criterion. EAs are very effective heuristics to solve combinatorial problems and are therefore of high relevance [127–129]. Instead of an exhaustive search over all possible band combinations, EAs provide strategies to optimally search the solution space and generate pseudo-optimal solutions.

The second algorithm represents a fundamentally different approach. Instead of estimating the information content of the spectral bands beforehand and trying to find

the best combination, embedded feature selection methods incorporate the selection of the most valuable features into a single training process of a specific machine learning algorithm. The algorithm optimises for its own purpose and in turn feeds back onto the input data and reveals information about the input features. This information can then be used to identify most relevant features and select those. The advantage of embedded algorithm lie in the fact that they can potentially be calculated very efficiently. The substantial advances in computational power over the last two decades have led to the resurgence of NNs and significant advances in deep learning algorithms and AEs were made. The feature selection capabilities of AEs are evaluated with respect to HSI data in particular. An algorithm developed in [62] aims at utilising AEs for embedded feature selection is modified and tailored to the specific needs of HSI data.

The remainder of this chapter is organised as follows. Section 4.2 introduces the above mentioned band selection algorithm utilising information theory and EAs. In Section 4.3, the AE based band selection approach is established and evaluated in detail. Where the two proposed algorithm are structurally quite different and yield their own conclusions, Section 4.4 discusses the two concepts from a wider angle and compares their advantages and drawbacks.

4.2 Fast band selection based on information theory and evolutionary algorithms

Based on the criteria established in Section 3.4, an unsupervised feature selection technique called MIMR-DGSA is proposed. The proposed algorithm is based on the work in [13]. It introduces the MIMR criterion, which helps identify pseudo-optimal band subsets by maximising the Shannon entropy of bands and minimising the MI between them. Details will be provided in Section 4.2.1. In [13] the usage of CSA to optimise the criterion is proposed. CSA is inspired by the human immune system and encodes solutions as antibodies. Each of the antibodies are evaluated for fitness by the MIMR criterion. Based on the degree of fitness, each antibody is cloned and the clones in turn mutated, where the number of clones is directly proportional to the fitness

and the amount of mutation inversely proportional, guaranteeing optimal exploration and exploitation of the search space (see Section 3.5 for a definition of exploration and exploitation). CSA however generates a huge amount of clones and can therefore get computationally relatively expensive. Additionally, it requires to tune six hyper-parameters to obtain best results. For these reasons, DGSA is developed here, which significantly reduces the number of solutions to be tested and has only three parameters to tune.

The MIMR criterion requires the calculation of the Shannon entropy and MI of hyperspectral bands, which in turn are based on estimates of the PDFs, and joint PDFs for the MI respectively, of the hyperspectral bands. To maximise the accuracy of the underlying PDF estimates, KDE techniques are employed. Both measures can be pre-calculated, as their number is finite and can be stored in lookup tables for quick evaluation. The naive approach of calculating the pairwise MI between all bands however quickly becomes very computationally expensive, especially for hyperspectral datasets that comprise a large amount of data and can last up to several days or weeks, according to the experiments conducted here. In [19], a fast algorithm for the pairwise calculation of MI of gene regulatory networks data is proposed. On the basis of this, VarBWFastMI is proposed, which significantly accelerates the calculation of MI and is also able to account for the varying shapes of density distributions in hyperspectral data. Based on VarBWFastMI, the discussed EAs are evaluated on three standard remote sensing HSI datasets and results are analysed with respect to performance, computational cost and reproducibility.

4.2.1 The Maximum Information Minimum Redundancy criterion using Clonal Selection Algorithm

For the MIMR-CSA algorithm proposed in [13], a criterion is defined for subset evaluation that maximises the information carried by the individual features whilst simultaneously minimising redundant information between all selected features of the subset. Let X with $|X| = s$ denote a set of band indices $x_i \in X$, where s denotes the number of desired bands. To maximise the information carried by the individual features, the

set of features X that minimises the the entropy of all wavelength images $\mathbf{I}_{x_i} \in \boldsymbol{I}$ of a hypercube $\boldsymbol{I}$ is defined by:

$$\arg\max_X \frac{1}{s} \sum_{i=1}^{s} H(\mathbf{I}_{x_i}) \tag{4.1}$$

Equally, to minimise the redundant information between the features, the X that minimises the MI of all its elements x_i is defined as:

$$\arg\min_X \frac{2}{s^2 - s} \sum_{i=1}^{s-1} \sum_{j=i+1}^{s} I(\mathbf{I}_{x_i}; \mathbf{I}_{x_j}) \tag{4.2}$$

In [13], Equations 4.1 and 4.2 are combined to the MIMR criterion as follows:

$$\arg\max_X \left(\sum_{i=1}^{s} H(\mathbf{I}_{x_i}) - \frac{2}{s-1} \sum_{i=1}^{s-1} \sum_{j=i+1}^{s} I(\mathbf{I}_{x_i}; \mathbf{I}_{x_j}) \right) \tag{4.3}$$

It should be noted that Equation 4.3 does not necessarily optimally minimise Equation 4.2 and maximise Equation 4.1 at the same time. The focus here however was not on redefining the criterion, but to establish a baseline for comparing EAs and their capability of optimising this criterion.

The higher the entropy of a feature, the higher its information. Equally, the lower the MI between two features, the lower the shared information, i.e. the redundancy between the features. This criterion measures intrinsic information of the features and does therefore not require any prior information. In the terminology of Section 3.4.1, this can be classified as an unsupervised subset evaluation with an independent criterion.

To optimise the MIMR criterion, CSA is proposed in [13]. CSA is an evolutionary algorithm inspired by the human immune system and models solutions as antibodies that react to an antigenic stimulus [130]. The affinity of the antibodies to antigens is defined by a fitness function, in this case the MIMR criterion. Antibodies then undergo a cloning an mutation process, where antibodies with a high affinity are cloned more often and these clones in turn mutated less. Antibodies with lower affinity spawn fewer clones which are mutated to a higher degree, to optimise their performance. A selection

process then selects the antibodies with the highest affinity and the whole process is repeated until the algorithm converges. The cloning and mutation are controlled by hyperparameters that require tuning. CSA has a total of six parameters (as detailed in Table 4.1), which make it relatively complex to use and the high number of clones leads to an equally high number of evaluations of the MIMR criterion. The proposed algorithm tries to address these issues.

4.2.2 Proposed approach

A flowchart of the proposed algorithm is outlined in Fig. 4.1, which has two main steps. The first step is the fast calculation of entropy and MI with VarBWFastMI and the second the MIMR-DGSA algorithm for band selection. Relevant details are presented in the following Sections.
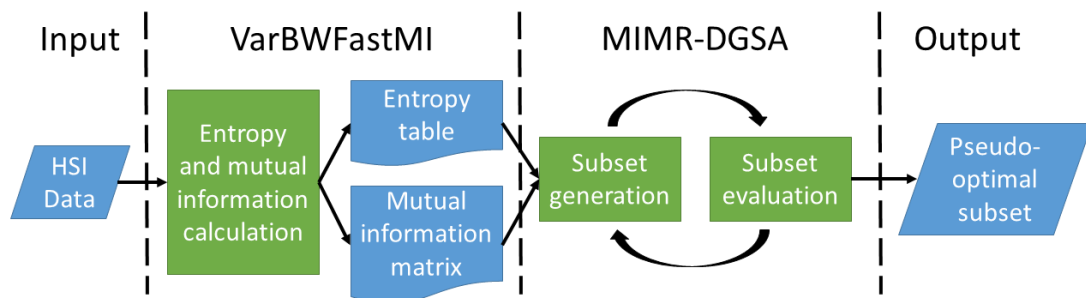
Figure 4.1: Procedure of the entire band selection algorithm. Mutual information and entropy are pre-calculated in lookup tables to speed up the MIMR evaluations.

Fast calculation of entropy and mutual information

The estimation of the PDFs of hyperspectral bands to calculate the entropy and MI pose some problems in practice that need to be addressed. Depending on the composition of the HSI image, each wavelength can potentially vary significantly in terms of grey value distribution. Figure 4.2 shows the histogram of two different wavelengths of the Indian Pines dataset. It is quite obvious that the two bands contain very different distributions of grey values and require therefore different bandwidths for the kernel density estimation. Here, a bandwidth estimation algorithm for Gaussian kernels based
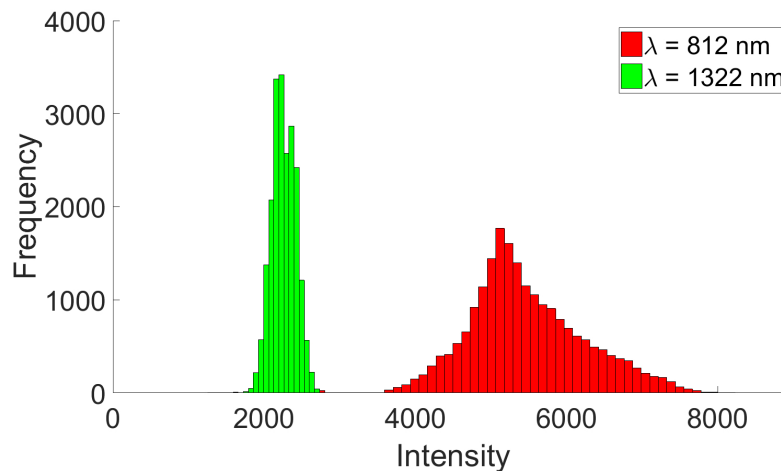
Figure 4.2: Histogram of wavelengths 812 and 1322 nm of the Indian Pines dataset. Both wavelength expose very different PDFs and therefore need individual bandwidths for density estimation.

on the principle of the Mean Integrated Squared Error (MISE) is used. Gaussian kernels are most commonly used and details about the bandwidth estimation can be found in [109]. The bandwidth estimation can deliver a pseudo optimal bandwidth for each hyperspectral band and therefore generate better estimates of the PDF and ultimately the entropy and MI. On the basis of the density estimate $\hat{p}_h$ with a given bandwidth h, an approximation of the entropy $H(X)$ of a band X with n sample points can be approximated by:

$$H(X) = \sum_{i=1}^{n} \hat{p}_h(x_i) \log \hat{p}_h(x_i) \tag{4.4}$$

The calculation of the MI encounters additional challenges. As seen in Equation 3.24, the MI requires the joint entropy of the two random variables, which in turn requires a joint density estimate. Using a Gaussian kernel function, the joint density can be estimated by Equation 3.28. The computational complexity however rises exponentially and for large datasets, the cost for the MI calculation becomes impractical. In [19], a fast algorithm for calculating the pairwise MI between features based on a Gaussian kernel density estimation is introduced for gene regulatory networks. The general idea proposed in [19] is to use the fact that the integral of Equation 3.24 can be approximated by the sample mean of the respective random variables. The implementa-

tion can then be adjusted to avoid repeated calculations of the same terms and therefore achieves a speedup. In this thesis, the existing algorithm is adjusted to VarBWFastMI by accounting for the fact that HSI bands require different kernel bandwidths for each band individually. Since different kernel bandwidths for each hyperspectral band are estimated, this needs to be considered for the pairwise MI calculation. Given the two bandwidths h_x and h_y for the two bands x and y, Equation 3.28 can be rewritten as:

$$\hat{p}(x,y) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{\sqrt{2\pi h_x h_y}} \times \exp\left(-\frac{1}{2}\left(\frac{(x-x_i)^2}{h_x^2} + \frac{(y-y_i)^2}{h_y^2}\right)\right) \tag{4.5}$$

The algorithm in [19] can simply be adapted to estimate the joint densities by Equation 4.5 to incorporate the variable bandwidths that are estimated. The implementation of VarBWFastMI is based on the Matlab implementation of the fast pairwise MI from [19] available in [131]. The code is altered to incorporate variable bandwidths.

Hyperspectral band selection using a modified discrete gravitational search algorithm

As a heuristic optimisation strategy developed in 2009 [132] that is inspired by Newton's gravitational laws, Gravitational Search Algorithm (GSA) has gained increasing popularity in recent years in various fields of optimisation [133, 134]. GSA interprets solutions as objects in an N-dimensional space that attract each other according to their mass and gravitational force, where higher masses represent better solutions. The basic concept is visualised in Figure 4.3a. For discrete problems, DGSA was developed in [20] based on GSA. In this thesis, DGSA is modified and adapted to the problem of hyperspectral band subset generation, by encoding band subsets as agents and defining a neighbourhood and movement concept. The modified DGSA can be defined as follows:

Let each band subset $X_i = (x_1, x_2, ..., x_N)$ be encoded by agents in an N-dimensional space, where N defines the number of desired bands. In each dimension d, the possible agent coordinates $x_d \in 1, 2, ..., M$ represent indices of all available bands $F = \{f_1, f_2, ..., f_M\}$. Duplicates are not allowed, i.e. $\forall i, j \in \{1, 2, ..., N, i \neq j\} \rightarrow x_i \neq x_j$

which means that no agent can have the same coordinate in more than one dimension. This is to avoid a subset containing the same band more than once. The representation of band subsets by agents in a search space is illustrated in Figure 4.3b. The mass M_i
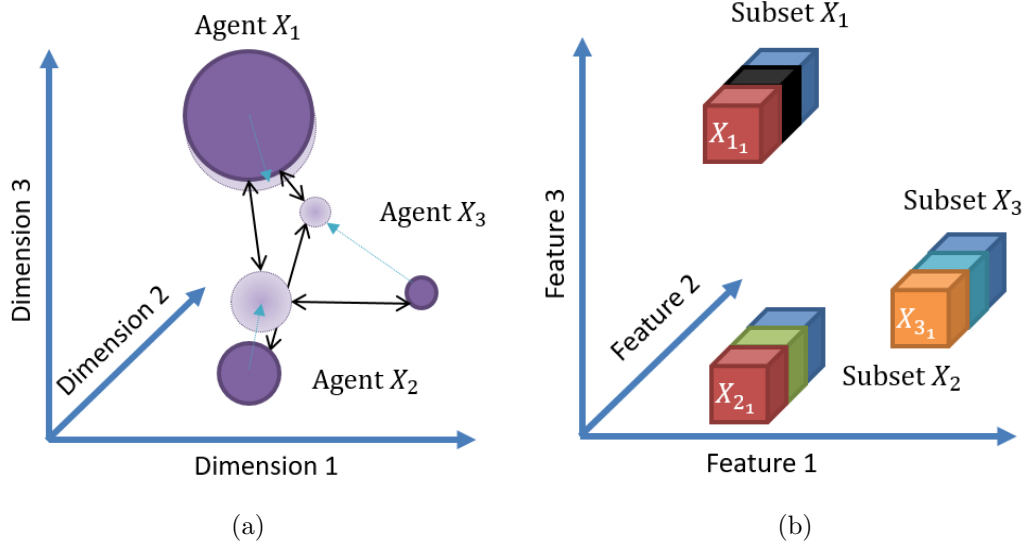
Figure 4.3: DGSA for three features. (a) Illustrates the agents and their interactions in the three dimensional space, and (b) the representation of band subsets through agents.

of agent i, the gravitational force F_{ij} and resulting acceleration a_{ij} from agent i to j as well as the normalised distance NR_{ij} are calculated according to [20]. M_i is defined by:

$$M_i = \frac{m_i}{\sum_{j=1}^{s} m_j}; \quad m_i = \frac{\text{fit}_i - \text{fit}_{worst}}{\text{fit}_{best} - \text{fit}_{worst}} \tag{4.6}$$

where s is the number of agents, fit_i represents the fitness of the i-th agent and fit_{best} and fit_{worst} the fitness of the best and the worst agent respectively. F_{ij} and a_{ij} can then be calculated by:

$$F_{ij} = \text{rand} \times G \times \frac{M_j M_i}{NR_{ij}} \times R_{ij} \tag{4.7}$$

$$a_{ij} = \left\lceil \frac{F_{ij}}{M_i} \right\rceil = \left\lceil \text{rand} \times G \times \frac{M_j}{NRij} \times R_{ij} \right\rceil \tag{4.8}$$

where G denotes the gravitational constant, R_{ij} the distance between agents i and j

and NR_{ij} the normalised distance in the interval $[0.5, 1]$, calculated by:

$$NR_{ij} = 0.5 + \frac{R_{ij}}{2R_{max}} \tag{4.9}$$

where R_{max} denotes the maximally possible distance between two agents.

The concept of distance between two band subsets is however redefined and inspired by the Manhattan distance. The distance between agents i and j can be interpreted as the number of edges that need to be passed along an undirected neighbourhood graph $N = (V, E)$ to reach j from i. N is a grid Graph where the set of nodes V contains all possible positions of agents and the edges E contain all connections between neighbouring nodes only. Neighbourhood is given when $\sum_{i=1}^{N} |x_i - y_i| = 1$, i.e. two agents only differ in one dimension by a value of exactly 1. One step can in this context be interpreted as moving agent i one edge along an N-dimensional grid towards agent j. Note that the movement is ambiguous as no preference as to which neighbour is selected is given in case of multiple paths between the two agents. Figure 4.4 illustrates movement in the Manhattan distance space between two agents for a two-dimensional search space. Two possible paths are indicated, but the number of paths is not limited to two. As a consequence, each of the possible paths is equally likely. One constraint here is that no band can be selected twice, which means that the number of possible neighbours is reduced. Paths are chosen randomly from the remaining paths to introduce a stochastic component and therefore prevent premature convergence.

The pseudo-code for the resulting MIMR-DGSA algorithm is provided in Algorithm 1 and can be summarised as follows:

A1.1 **Input** The algorithm expects the number of desired bands to be selected n_b. The parameters that influence the performance are the number of iterations $iter_{max}$, the population size s and the number of candidates for the initialisation n_c

A1.2 **Initialisation** The initial number of agents K that are used to calculate is set to all agents. This number is linearly reduced to 1. The gravitational constant G is set to 1 and linearly reduced to 0. The initial population P of agents is
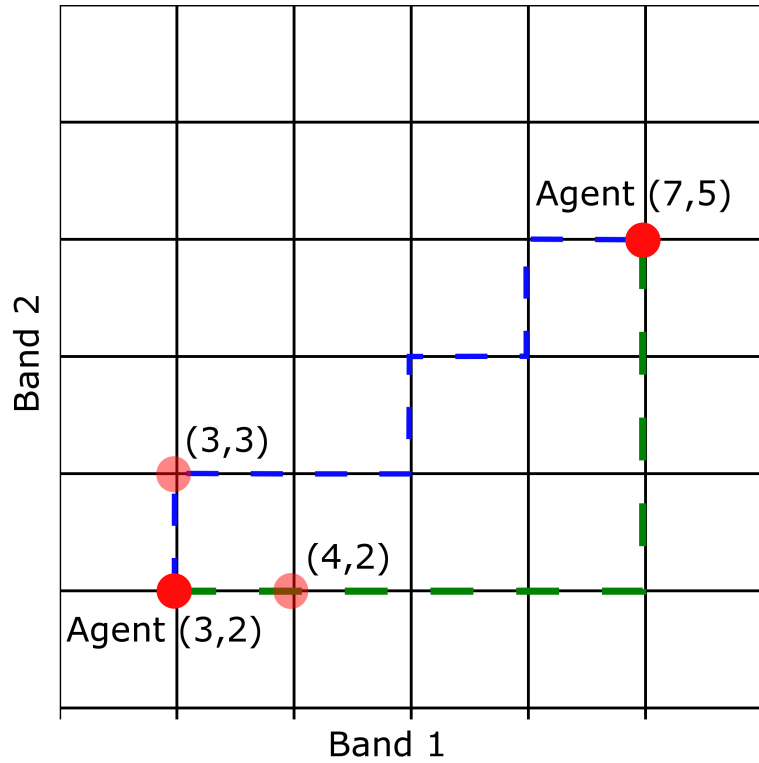
Figure 4.4: Illustration of movement in the Manhattan distance space in a two dimensional space. The agent at coordinates (3,2) is moving towards the agent at (7,5) and has two options at each step. This creates a number of possible shortest paths which are all equally likely. The blue path to (3,3) does however not satisfy the condition that no band can be selected twice. The step to (4,2) will be taken and from there, a number of new paths are possible.

generated by a randomised greedy initialisation where n_c random bands out of all are selected as candidates and the one that produces the highest fitness is chosen and added to a set until the set reaches the desired size n_b.

A1.3 **Main Loop** In the main loop, both K and the gravitational constant G are reduced by a linear reduction function. K should equal 1 in the last iteration. K_{best} contains the set of all agents that exert their gravitational force. As suggested in [20], the K_{best} set contains the globally K best agents out of all iterations instead of the local K best agents of the current iteration. The mass and acceleration of each agent are calculated according to Equations 4.6 and 4.8 respectively. The loop terminates after $iter_{max}$ iterations.

A1.4 **Movement operator** The movement operator Move(.) takes an agent X_i of the current population $P(iter)$ as an input, along with the K_{best} agents and the previously calculated accelerations a_{ij} from the agent i to all agents j in K_{best}. Using these accelerations, the agent i is moved towards all agents j as illustrated in Figure 4.4 successively. It needs to be specified in which order the K_{best} agents exert their force onto other agents. As stated in [20], later movements have a more significant impact on the quality of the solution, which is why the priority is calculated by the inverse mass. This procedure is done for each agent in the population.

A1.5 **Local search** At the end of each iteration, a local search is performed for each agent. This is based on the Hill climbing algorithm [135]. The worst performing band of the current subset is replaced by the best performing band of all remaining bands, defined by the maximum entropy. The search terminates when no neighbour can further improve the fitness of the subset. The algorithm terminates after $iter_{max}$ iterations and the current best solution poses the pseudo-optimal subset of the selected bands.

4.2.3 Experimental results

For performance assessment, the proposed MIMR-DGSA algorithm was tested on three different hyperspectral remote sensing datasets; the Indian Pines (Appendix A.1), the Salinas (Appendix A.2) and the Pavia University dataset (Appendix A.3).

Experimental setup

To evaluate the performance of the proposed MIMR-DGSA algorithm, it was compared with the original CSA version as well as PSO and FA. The individual parameter settings are listed in Table 4.1. For CSA, the settings are based on [13], whereas for PSO and FA, the parameters of the respective literature were used as a basis and were empirically adjusted for optimal results on the data used in this paper. All algorithms are compared with respect to their band selection capabilities and time consumption.

Algorithm 1 MIMR-DGSA Algorithm

1: **Input:** n_b: Number of desired bands; s: Size of population or number of agents; $iter_{max}$: Maximum number of iterations; n_c: Number of candidates for generation of initial population // See A1.1

2: **Output:** Best solution X_{best} found;

3: ▷ Initialisation See A1.2

4: $iter \leftarrow 0$; $K \leftarrow s$; $G_{start} \leftarrow 1$; $G_{end} \leftarrow 0$

5: Generate initial population $P(iter)$ with s agents of length n_b

6: Evaluate fitness $fit_i = \text{MIMR}(X_i)$ and mass m_i of each agent $X_i \in P(iter)$ // See Equation 4.6

7: ▷ Main loop See A1.3

8: **while** $iter < iter_{max}$ **do**

9:      Update G, K by linear reduction functions

10:      Update K_{best} by selecting global K best solutions

11:      Calculate acceleration a_{ij} for each agent $X_i \in P(iter)$ with respect to each $X_j \in Kbest$ // See Equation 4.8

12:      **for** $i = 1$ to s **do**

13:        $X_i \leftarrow \text{Move}(X_i, K, K_{best}, a_{ij})$ // See A1.4

14:      **end for**

15:      **for** $i = 1$ to s **do**

16:        $X_i \leftarrow \text{LocalSearch}(X_i)$ // See A1.5

17:      **end for**

18:      $P(iter + 1) \leftarrow P(iter)$

19:      Evaluate fitness $fit_i = \text{MIMR}(X_i)$ and mass m_i of each agent $X_i \in P(iter + 1)$

20:      Store best solution $X_{best} \in P(iter + 1) \cup K_{best}$

21:      $iter \leftarrow iter + 1$

22: **end while**

The band selection capabilities were assessed by the pixel-wise classification. An SVM with Gaussian RBF kernel whose parameters C and γ were tuned by a grid search, i.e. selecting 20% of the pixels of each dataset's classes randomly for training and the remaining 80% for validation. The test- and validation-set splitting was repeated 10 times for each dataset and 3 runs of each algorithm were performed for each set making it 30 runs per dataset. As a state-of-the-art unsupervised feature selection benchmark for the classification performance, the Ward's Linkage strategy using Mutual Information (WaLuMI) [136] algorithm was applied. It hierarchically groups the spectral bands by a distance measure based on MI and selects a representative of each group as the band subset. WaLuMI was chosen as it performs best among all compared algorithms in [13] and therefore serves as a baseline. To compare the performance of the MIMR criterion, the Fuzzy C-Means Clustering (FCM) [137] method was applied in combination with CSA and ultimately, a classification using all bands of each dataset was compared with that of the selected features.

Entropy and mutual information

For each of the three datasets, the lookup tables for the entropy and MI were calculated with the proposed VarBWFastMI algorithm. Results are visualised in Figures 4.11 and 4.12 along the class mean spectra of each class in all three datasets. As expected, the Indian Pines and Salinas datasets show a strong structural similarity for both the entropy and MI as they are captured with the same sensor and contain similar vegetation scenes. The bands on the edges of the water absorption regions in the Indian Pines and Salinas datasets show a very low entropy and a low MI with the rest of the bands. Wavelength number #40 - #100 seem to carry the most information as the entropy is the highest in that range. The Pavia University in contrast contains more manmade objects and covers only a portion of the spectral range of the other two datasets and is therefore quite different. It has a remarkably uniform high entropy over almost all bands. In bandnumbers #70 - #80, the entropy decreases slightly as the class spectra seem to be less distributed and closer together. The MI equally shows a

Table 4.1: Parameter configurations for the different algorithms

CSA			DGSA		
Population size s	50		Population size s	30	
Maximum iterations $iter_{max}$	100		Maximum iterations $iter_{max}$	30	
Displaced antibodies d	5		Candidates for initialisation n_c	10	
Number of clones n_{cl}	2500				
Mutation probability n_m	5				
Selection probability n_s	0.5				
PSO			FA		
Population size s	25		Population size s	10	
Maximum iterations $iter_{max}$	100		Maximum iterations $iter_{max}$	100	
Acceleration coefficient 1 c_1	2		Step size factor α	0.5	
Acceleration coefficient 2 c_2	2		Maximum attraction β_0	0.2	
Range of inertia weight w	0.9		Absorption coefficient γ	1	

decline as most bands seem to undergo a distinct change specific to their class.

The speed-up achieved by the VarBWFastMI algorithm compared to the naive approach is compared in Tables 4.2 and 4.3. The naive approach entails estimating the univariate and bivariate PDFs of each band and band combinations individually, while for the MI matrix, only the upper triangle needs to be calculated and the lower one can be mirrored. The kernel density estimation is done with the Matlab KDE Toolbox available at [138]. It employs kd-trees to faster estimate the kernel density. The VarBWFastMI algorithm was implemented in Matlab as well and both algorithms were run on an Intel Core i5 CPU at 3.20 GHz with 16 GB RAM and were performed on the Indian Pines dataset. In Table 4.2, different numbers of bands out of the 200 available ones were selected randomly in 5 runs and the mean runtime was measured. In Table 4.3, different numbers of samples were used for the calculation. These numbers stem from taking 1%, 5%, 10%, 25% and 50% of the samples of each class. One can see that with increasing number of bands, the speed-up also increases whereas the speed-up

seems to be not directly correlated to the numbers of samples. As each band has a very different PDF, with very different numbers of samples and distribution over the intensity range, the time consumption cannot be linearly scaled up with increasing numbers of samples. A definite speed-up factor cannot be established but it is evident that the proposed VarBWFastMI performs much faster than the naive approach by a large factor. Especially for large datasets such as the Salinas dataset, this can significantly reduce the computational cost from several weeks to only days.

Table 4.2: Time consumption in seconds for the MI with varying numbers of bands on the Indian Pines dataset with a fixed number of 1031 samples

# Bands	10	20	50	75	100	150
Naive	53.4	303.1	1591.7	3608	6532.1	14149
VarBWFastMI	5.7	9.3	30.0	47.1	67.1	116.5
Speed-up	9.3	32.6	53.1	76.6	97.5	122.0

Table 4.3: Time consumption in seconds for the MI with varying numbers of samples for the Indian Pines dataset with all 200 bands

# Samples	110	520	1031	2569	5128
Naive	33.1	379.4	1318.1	7249.9	25188
VarBWFastMI	0.6	2.3	7.9	43.7	169.6
Speed-up	55.1	165.0	166.8	165.9	148.5

Values obtained by the naive approach and VarBWFastMI were compared and differences are at a magnitude of around 10×10^{-5}. i.e. both approaches yield almost identical results.

DGSA Parameter analysis

All parameters were plotted against the evaluation of the MIMR fitness function for the Indian Pines dataset.
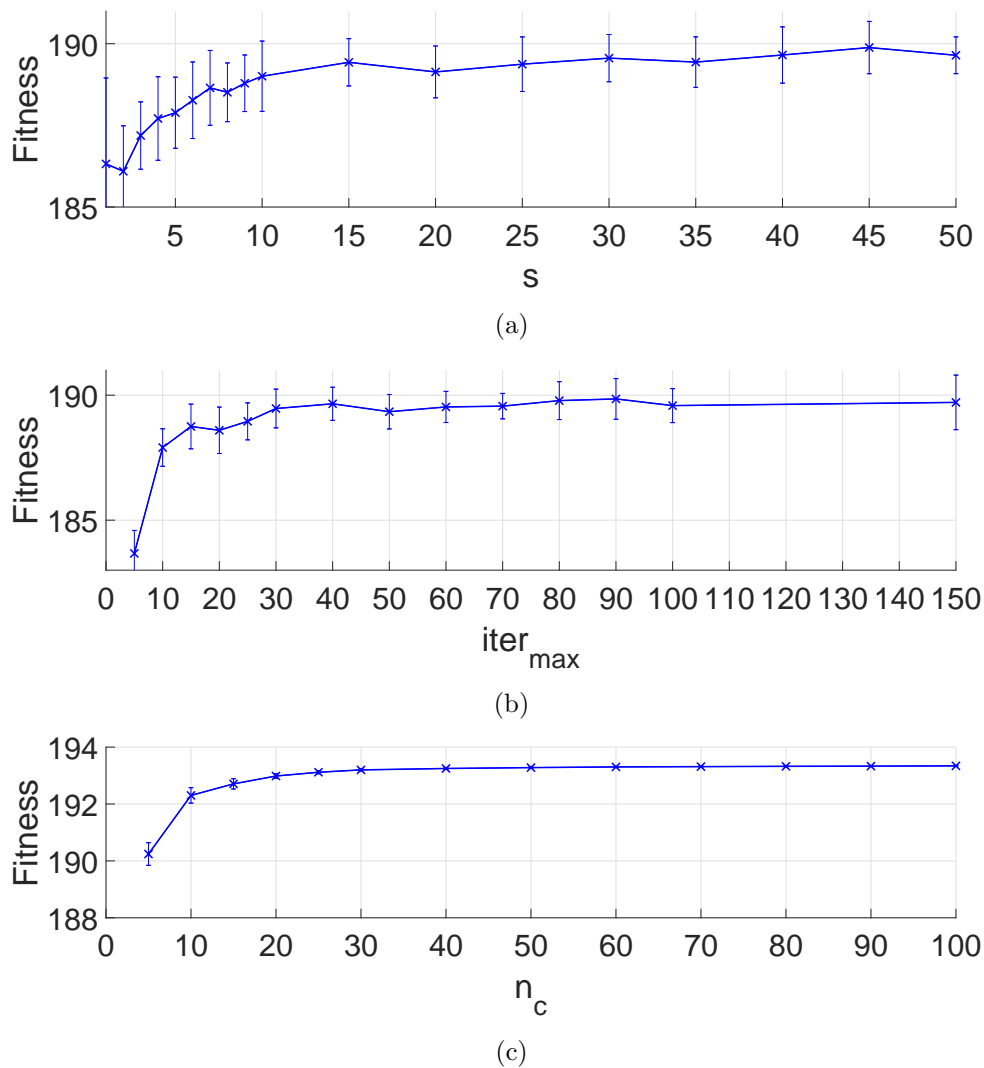
Figure 4.5: Sensitivity of DGSA parameters to the optimisation performance on the Indian Pines dataset. (a) Population size, (b) number of iterations, and (c) number of candidates.
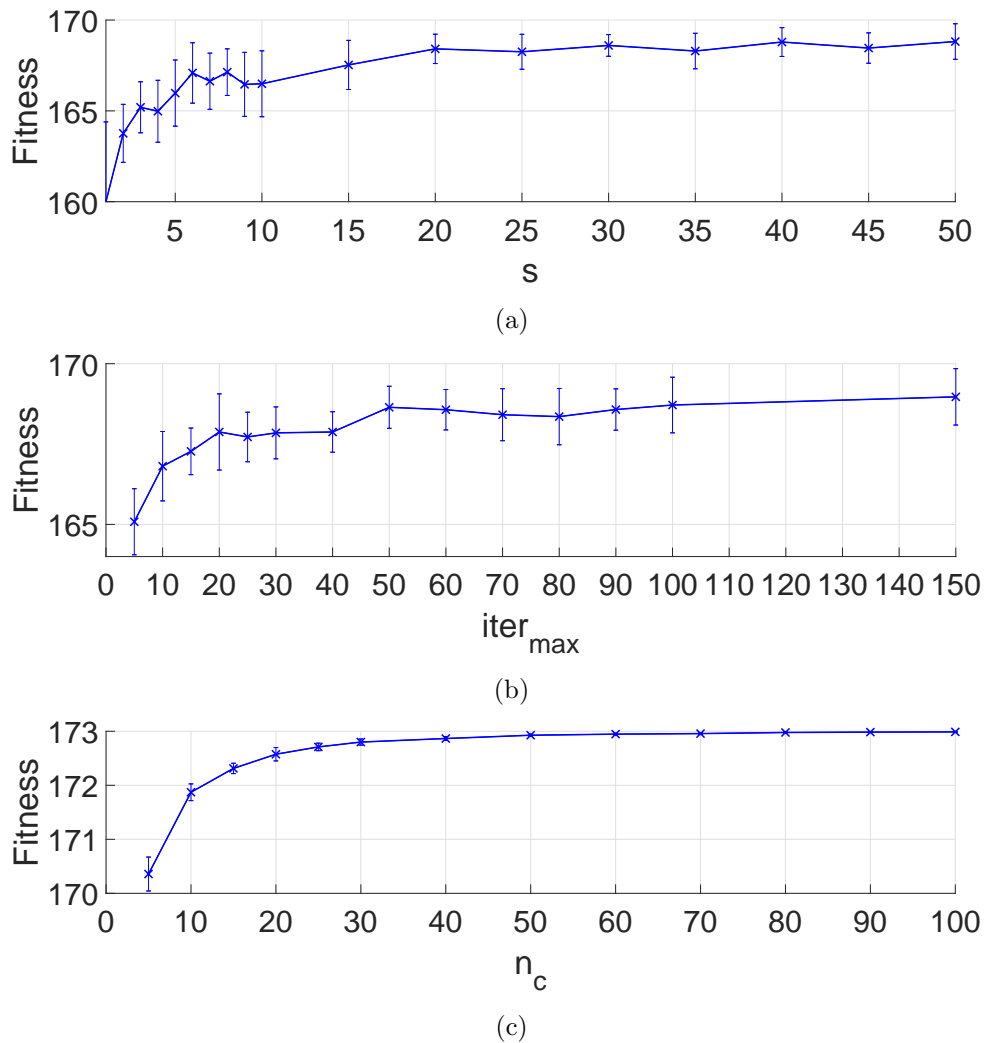
Figure 4.6: Sensitivity of DGSA parameters to the optimisation performance on the Salinas dataset. (a) Population size, (b) number of iterations, and (c) number of candidates.

The population size s and the number of iterations $iter_{max}$ are the two key parameters influencing the performance of DGSA. Both were analysed in Figures 4.5a and 4.5b for a fixed number of 30 bands on the Indian Pines dataset and in Figures 4.6a and 4.6b for the Salinas dataset. To magnify the impact of both parameters, DGSA was performed with random initialisation. As expected, both parameters increase the performance by increasing their values. They are chosen to be as minimal as possible to achieve maximum optimisation capacity with minimal computational effort. Based on

these results, s and $iter_{max}$ are both set to 30 in future experiments for MIMR-DGSA for all datasets to generate the same conditions.

The number of candidates n_c for the initialisation was analysed in Figure 4.5c. By increasing this number, the chances of picking a good solution increase as well. However, to guarantee a good exploration of the search space, a trade-off between subset fitness and population diversity is mandatory. Therefore, n_c is suggested to be set to 10.

Computational complexity and runtime analysis

As stated in [56], PSO has a computational complexity of $\mathcal{O}(i \times s)$ which is linearly correlated to the number of iterations i and the population size s. Due to the interaction between the fireflies, FA is quadratically correlated to the size of the population and has a complexity of $\mathcal{O}(i \times s^2)$. As stated in [13], CSA has a time complexity of $\mathcal{O}(i \times n_{cl})$, which is linearly dependent on the number of clones n_{cl} and iterations i. DGSA has a complexity of $\mathcal{O}(i \times s \times K)$ which is additionally linearly correlated to the K_{best} solutions. All algorithms however share a quadratic runtime of the MIMR evaluations $\mathcal{O}(n_b^2)$, which is quadratically dependent on the number of selected bands n_b. In comparison to CSA, the number of MIMR evaluations per iteration in DGSA is limited to the number of initial agents s and the decreasing number of K best solutions, whereas CSA generates a huge number of clones in every iteration, leading to a significantly longer runtime. The initialisation of DGSA has a time complexity of $\mathcal{O}(s \times n_c \times n_b^2)$, which is linearly correlated to the number of candidates n_c and quadratically correlated to n_b. All algorithms share the dependence on number of iteration and the quadratic runtime of the MIMR evaluation, showing the importance of pre-calculating the entropy and MI. The main differences are rooted in the population size and the interaction between the solutions. DGSA has an additional initialisation step which can potentially decrease the efficiency. In this, case, the number of candidates for the initialisation is relatively low, which is why this step does not have a big effect on runtime.

To compare the time consumption, all algorithms were performed on the Indian Pines dataset with different numbers of selected bands. For consistency, both CSA and DGSA algorithms were performed with the same population size and are terminated
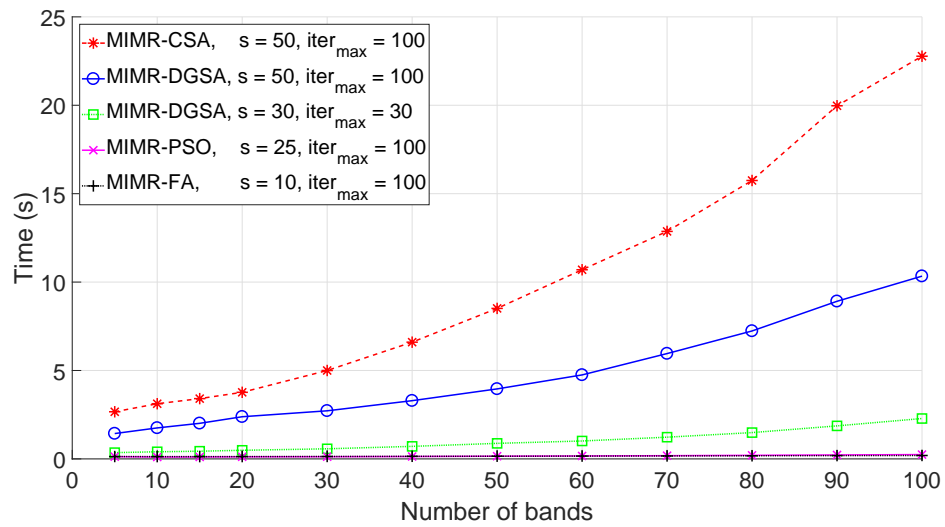
(a)
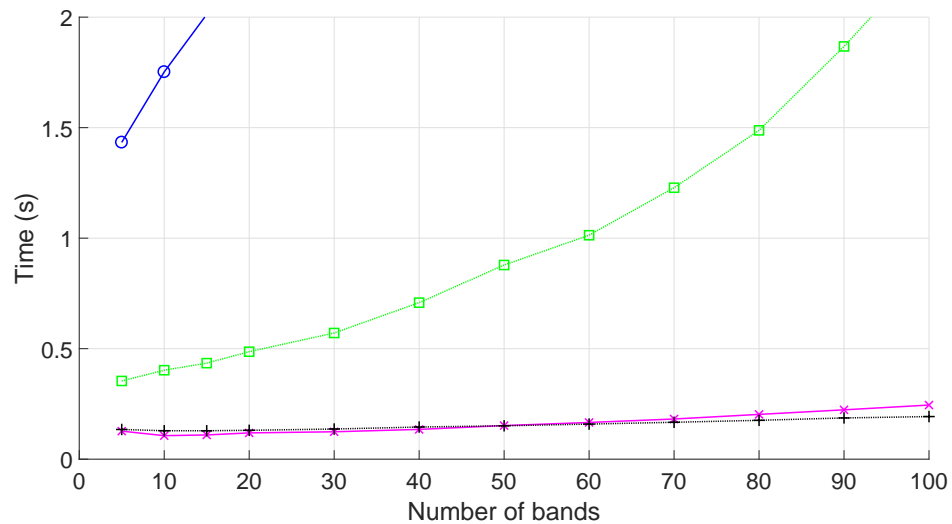
(b)

Figure 4.7: Time consumption of band selection algorithms with different parameter settings. s denotes the population size and $iter_{max}$ the maximum number of iterations. (a) Shows all five algorithms and (b) a closeup of MIMR-DGSA, MIMR-PSO and MIMR-FA with different numbers of selected bands.

after the same number of iterations. Figure 4.5 shows that increasing to number of iterations does not increase the optimisation performance significantly. Additionally, MIMR-DGSA was performed with optimised population size and iterations, as established in Section 4.2.3. The number of candidates for the initialisation of DGSA is set

to 10. For PSO and FA, the population size and iterations were set according to [56] and [57] respectively and slightly adjusted based on empirical values and all other parameters are set as specified in Table 4.1. As seen from the time measurements in Figure 4.7, MIMR-DGSA with the same s and $iter_{max}$ as MIMR-CSA performs about twice as fast. With the parameter settings established in Section 4.2.3, MIMR-DGSA only requires a fraction of the time of MIMR-CSA, where for 150 features, MIMR-DGSA takes only 5 seconds compared to over 50 seconds for MIMR-CSA. The increased time consumption of CSA is rooted in the relatively high number of clones per antigen, whereas DGSA only has a limited number of agents with a decreasing number of K best agents. PSO and FA both perform very similar and both outperform CSA and DGSA due to their straightforward implementation of the movement strategy. DGSA suffers in this respect due to the elaborate neighbourhood and movement concept.

Classification performance

The classification accuracies using the selected features are compared to evaluate the efficacy of the band selection approaches. MIMR-FA, MIMR-PSO, MIMR-CSA, MIMR-DGSA as well as WaLuMI were performed selecting 30 bands on the Indian Pines and Salinas datasets and 20 bands on the Pavia University dataset. The OA, AA and Kappa coefficient were calculated in every case alongside the individual class accuracies. Results are summarised in Tables 4.4, 4.5 and 4.6 for comparison.

As seen in Table 4.4, in terms of OA, MIMR-PSO performs best for 30 bands on the Indian Pines dataset, where MIMR-CSA and MIMR-DGSA perform roughly similar. Only FCM-DGSA performs significantly worse. As seen in Figure 4.8, MIMR-DGSA and MIMR-CSA also perform very similar for different numbers of selected bands, outperforming MIMR-FA and WaLuMI, whereas MIMR-PSO performs best selecting 90 or less bands. For the Salinas dataset, MIMR-FA performs best as seen in Table 4.5 and Figure 4.9. Again MIMR-CSA and MIMR-DGSA show a very similar performance for different numbers of bands but are outperformed by WaLuMI and notably FCM-DGSA when the number of bands is low. For the Pavia University dataset, MIMR in combination with CSA, DGSA, FA and PSO all outperform WaLuMI and FCM-DGSA

67

Table 4.4: Comparison of classification results selecting 30 bands for different algorithms on the Indian Pines dataset

Class	MIMR-CSA	MIMR-DGSA	WaLuMI	FCM-DGSA	MIMR-PSO	MIMR-FA
1	72.2±9.9	75.7±9.7	71.7±18	63.8±13	**79.0±8.1**	73.3±12
2	76.9±3.0	76.3±3.6	**79.0±2.1**	64.1±5.4	78.0±2.9	72.9±5.9
3	72.5±3.0	72.0±3.6	69.7±3.8	57.2±5.9	**73.3±3.9**	68.6±7.0
4	67.1±4.7	68.9±5.3	**69.1±3.2**	51.2±7.2	67.4±4.4	64.0±8.7
5	**92.8±1.1**	92.7±0.9	91.2±1.1	88.1±2.3	91.6±2.2	91.1±3.0
6	98.0±1.3	**98.0±0.8**	97.0±1.4	95.1±1.8	96.5±1.8	96.2±1.9
7	88.2±2.3	86.7±5.3	**87.3±2.0**	60.0±16	84.8±9.0	80.0±10
8	**98.4±0.8**	98.0±1.4	97.7±1.1	96.6±2.0	96.9±1.6	97.0±1.9
9	63.3±17	63.8±19	**70.0±18**	22.3±15	51.0±24	56.3±22
10	77.8±3.2	77.0±3.3	76.7±3.1	63.4±3.5	**80.0±3.3**	71.7±10
11	**85.3±1.3**	84.4±1.4	83.6±2.1	78.0±2.0	85.3±1.5	82.7±2.4
12	73.4±2.9	75.6±3.8	73.6±4.9	50.1±6.9	**81.3±3.6**	79.7±3.6
13	96.2±1.9	96.3±2.1	97.2±2.2	92.0±4.7	97.3±1.5	**98.0±1.2**
14	95.8±0.9	95.7±0.9	95.1±1.1	94.4±1.4	**95.9±1.1**	95.5±1.2
15	59.6±3.3	60.3±2.7	56.9±2.9	47.8±5.6	**61.8±4.7**	59.5±5.6
16	86.1±6.8	85.9±9.0	87.3±6.8	89.1±4.3	**91.7±5.1**	88.9±6.1
OA	83.6±0.8	83.4±0.8	82.9±0.5	74.4±2.0	**84.4±1.1**	81.5±2.9
AA	81.5±1.3	81.7±1.6	81.4±1.2	69.6±2.7	**82.0±1.8**	79.7±3.4
Kappa	81.3±1.0	81.0±0.9	80.5±0.5	70.6±2.3	**82.2±1.3**	78.8±3.4

Table 4.5: Comparison of classification results selecting 30 bands for different algorithms on the Salinas dataset

Class	MIMR-CSA	MIMR-DGSA	WaLuMI	FCM-DGSA	MIMR-PSO	MIMR-FA
1	**99.5±0.2**	**99.5±0.2**	99.4±0.2	99.5±0.4	99.4±0.4	99.4±0.4
2	99.8±0.2	99.8±0.2	**99.8±0.1**	**99.8±0.1**	**99.8±0.1**	99.7±0.2
3	98.6±0.4	98.8±0.4	**99.6±0.2**	99.3±0.4	99.3±0.4	99.2±0.5
4	99.4±0.2	99.4±0.3	**99.5±0.2**	99.2±0.4	99.3±0.4	99.4±0.4
5	98.2±0.4	98.4±0.6	**99.0±0.4**	98.9±0.5	98.9±0.4	98.8±0.6
6	99.8±0.1	99.8±0.1	**99.9±0.1**	**99.9±0.1**	99.8±0.1	99.8±0.1
7	**99.8±0.1**	**99.8±0.1**	99.7±0.1	99.6±0.2	99.6±0.2	99.6±0.2
8	87.5±0.8	88.3±0.9	88.6±0.5	88.8±0.6	88.8±0.7	**89.2±0.6**
9	98.4±0.4	99.2±0.7	**99.8±0.2**	99.7±0.1	**99.8±0.1**	**99.8±0.1**
10	94.8±0.6	95.6±0.8	**97.3±0.6**	96.6±0.9	97.2±0.6	**97.3±0.6**
11	92.7±1.6	96.1±2.5	98.8±0.4	98.3±0.9	98.4±1.0	**99.0±0.7**
12	99.8±0.2	99.8±0.2	99.8±0.1	**99.9±0.1**	99.8±0.4	**99.9±0.1**
13	99.4±0.3	**99.5±0.3**	98.9±0.6	99.1±0.6	99.2±0.5	99.2±0.5
14	97.7±1.0	97.9±0.9	97.1±1.0	97.4±1.3	97.4±1.3	**98.0±1.2**
15	71.5±1.4	74.3±2.3	73.6±1.1	75.2±1.3	74.6±2.6	**75.8±2.2**
16	**99.1±0.2**	**99.1±0.2**	98.9±0.2	98.8±0.3	98.8±0.3	98.9±0.3
OA	92.6±0.3	93.4±0.5	93.6±0.1	93.8±0.2	93.8±0.4	**94.0±0.4**
AA	96.0±0.2	96.6±0.4	96.9±0.1	96.9±0.2	96.9±0.3	**97.1±0.3**
Kappa	91.8±0.3	92.6±0.6	92.9±0.1	93.1±0.3	93.0±0.5	**93.3±0.4**

Table 4.6: Comparison of classification results selecting 20 bands for different algorithms on the Pavia University dataset

Class	MIMR-CSA	MIMR-DGSA	WaLuMI	FCM-DGSA	MIMR-PSO	MIMR-FA
1	92.2±0.6	92.1±1.0	92.9±0.4	90.3±1.2	91.7±1.4	**93.5±0.8**
2	**97.9±0.2**	97.7±0.2	96.9±0.2	96.5±0.5	97.0±0.5	97.1±0.5
3	75.3±1.7	74.7±3.7	76.0±1.0	66.8±6.9	74.9±5.1	**77.4±3.3**
4	**94.1±0.8**	93.8±0.9	91.5±0.7	90.4±1.1	92.3±1.2	92.7±1.4
5	99.3±0.3	**99.4±0.3**	99.3±0.2	98.5±0.6	99.1±0.3	99.2±0.3
6	**89.6±0.6**	88.4±1.1	81.1±1.1	64.6±9.9	77.4±9.0	83.0±5.5
7	78.7±1.8	78.0±3.4	80.9±1.7	80.2±2.1	82.2±1.9	**83.8±1.7**
8	87.4±0.9	87.6±1.6	87.9±0.7	86.8±1.8	88.7±1.6	**89.8±1.1**
9	99.9±0.2	**99.9±0.1**	99.8±0.2	99.7±0.2	99.8±0.1	99.8±0.2
OA	**93.2±0.1**	92.9±0.5	91.9±0.1	88.7±1.8	91.4±1.7	92.7±1.1
AA	90.5±0.3	90.2±0.9	89.6±0.2	86.0±2.2	89.2±1.9	**90.7±1.2**
Kappa	**91.0±0.2**	90.6±0.7	89.2±0.1	84.8±2.5	88.5±2.3	90.3±1.5

Table 4.7: Mean OA, AA and Kappa coefficient over the three datasets of the different algorithms

	MIMR-CSA	MIMR-DGSA	WaLuMI	FCM-DGSA	MIMR-PSO	MIMR-FA
OA	89.80±0.40	**89.90±0.60**	89.47±0.23	85.63±1.33	89.87±1.07	89.40±1.47
AA	89.33±0.60	**89.50±0.97**	89.30±0.50	84.17±1.70	89.37±1.33	89.17±1.63
Kappa	88.03±0.50	**88.07±0.73**	87.53±0.23	82.83±1.70	87.90±1.37	87.47±1.77

for any number of bands up to 75, as seen in Figure 4.10 where CSA seems to perform best when selecting 20 bands. Both MIMR-DGSA and MIMR-CSA again perform very similar as shown in Table 4.6 and Figure 4.10. Another observation that can be made specifically from Tables 4.4 and 4.6 is that MIMR-PSO and MIMR-FA yield a larger standard deviation. This might hint at the fact that they have less optimisation qualities for the MIMR criterion and the selected bands are more random, which is examined in Section 4.2.3. Even though this leads to a higher mean accuracy, the algorithms are less reliable in individual runs. This leads to the overall conclusion that FA and PSO can potentially achieve a higher classification accuracy, whereas CSA and DGSA perform slightly less but more robust. None of the presented algorithms consistently indicates optimal performance. For the individual class accuracies, very high standard deviations can be observed in some cases. This is likely due to the unevenly spread number of samples for each class. Class 9 (Oats) for example shows very high standard deviation for all algorithms, and in Appendix A.1, one can see that the class only has 20 available samples.

To investigate this, the average measures of the OA, AA and Kappa coefficient over the three datasets were compared, as shown in Table 4.7. MIMR-DGSA seems to slightly outperform all other algorithms. This hints at the fact that DGSA performs better in terms of generalising over different datasets, but in individual cases, might perform inferior. Hence, even though FA and PSO may produce higher classification accuracy in individual cases, their overall performance for band selection can be compromised due to inferior generalisation capabilities. This is evaluated in detail in the following section.

Optimisation performance

The optimisation performance of the EAs with respect to the MIMR criterion is investigated. The objective is to maximise the individual entropies of selected bands and minimise the associated MI. A higher MIMR value indicates a better subset. The actual value is dependent on the constitution of each dataset, but algorithms can be compared among each other within one dataset. Table 4.8 summarises the mean MIMR

71
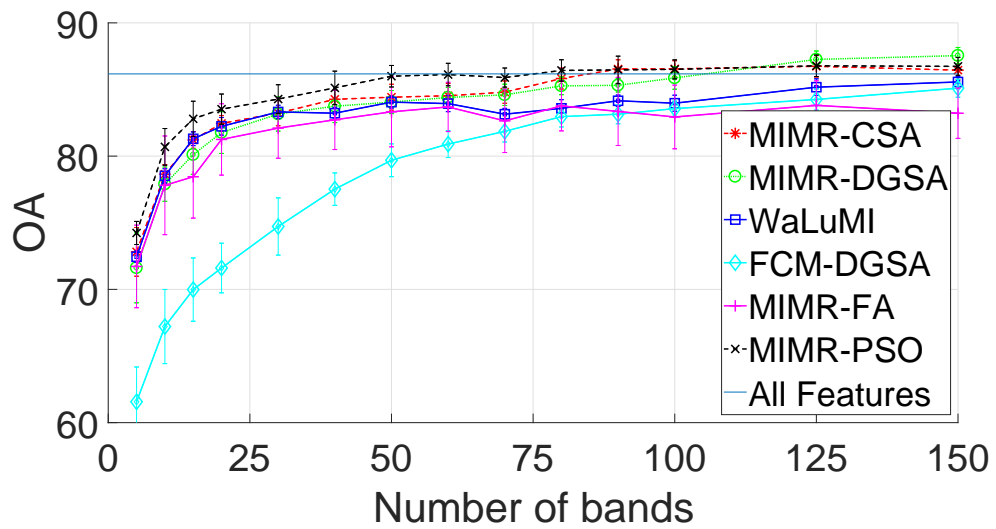
Figure 4.8: Performance comparison of all band selection algorithms on the Indian Pines dataset. The horizontal line indicates the classification accuracy achieved using all original 200 bands without prior band selection.
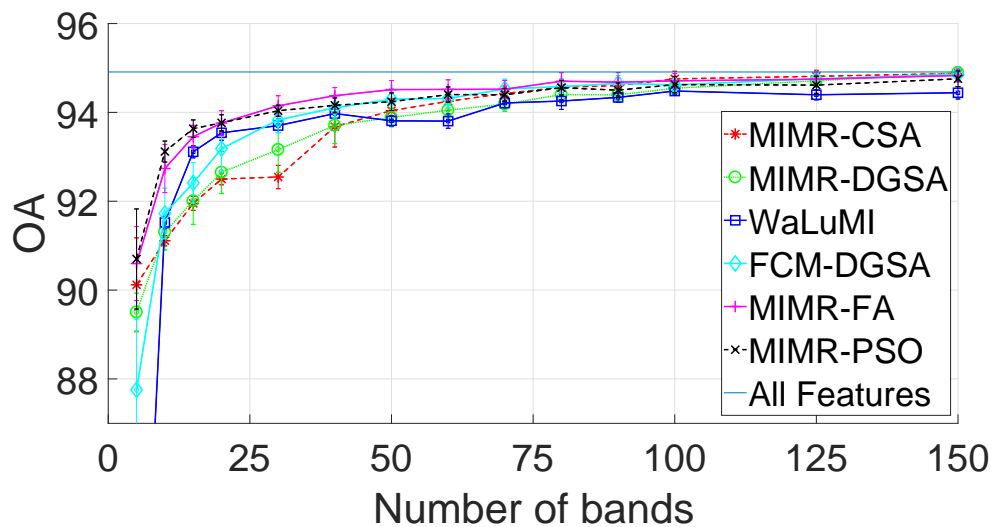
Figure 4.9: Performance comparison of all band selection algorithms on the Salinas dataset. The horizontal line indicates the classification accuracy achieved using all original 200 bands without prior band selection.

value achieved after the last iteration of the four EAs of 20 runs. One can see that FA performs worst on all datasets, whereas CSA and DGSA perform very similar. PSO performs better than FA and even becomes even with DGSA for the Pavia University
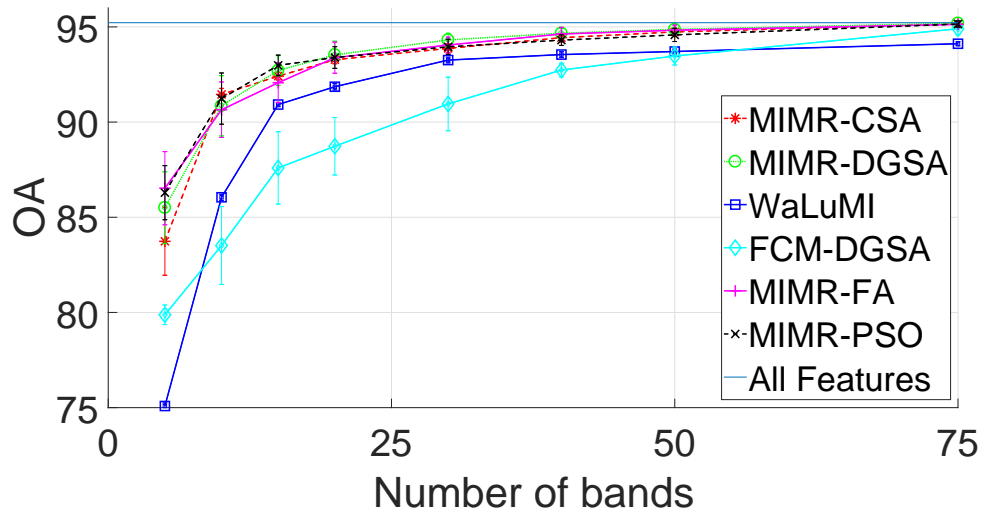
Figure 4.10: Performance comparison of all band selection algorithms on the Pavia University dataset. The horizontal line indicates the classification accuracy achieved using all original 103 bands without prior band selection.

Table 4.8: Comparison of optimisation performance of the different optimisation algorithms. Values indicate the evaluation of the MIMR criterion of the best solution achieved by each optimisation algorithm

Dataset	# Bands	FA	PSO	CSA	DGSA
Indian Pines	30	175	188	**193**	**193**
Salinas	30	159	168	**173**	172
Pavia University	20	138	**141**	**141**	**141**

dataset but worse for the other datasets. As seen in Figure 4.12, the MI and especially the entropy of most bands in the Pavia University dataset are very similar which might suggest that most band combinations yield similar MIMR values. This explains why the optimisation performance of all algorithms is very similar. These findings confirm the above assumption that optimising the MIMR criterion does not necessarily imply an optimal classification accuracy. However, better optimisation of the MIMR criterion results in a more robust accuracy and therefore band selection performance. The fact that FA and PSO yield a better classification accuracy with an inferior optimisation capability of the chosen MIMR criterion hence exposes shortcomings of the criterion

rather than a superior band selection quality of the optimisation algorithm. For this reason, both CSA and DGSA are considered as more suitable for the proposed task.

Validation of selected bands
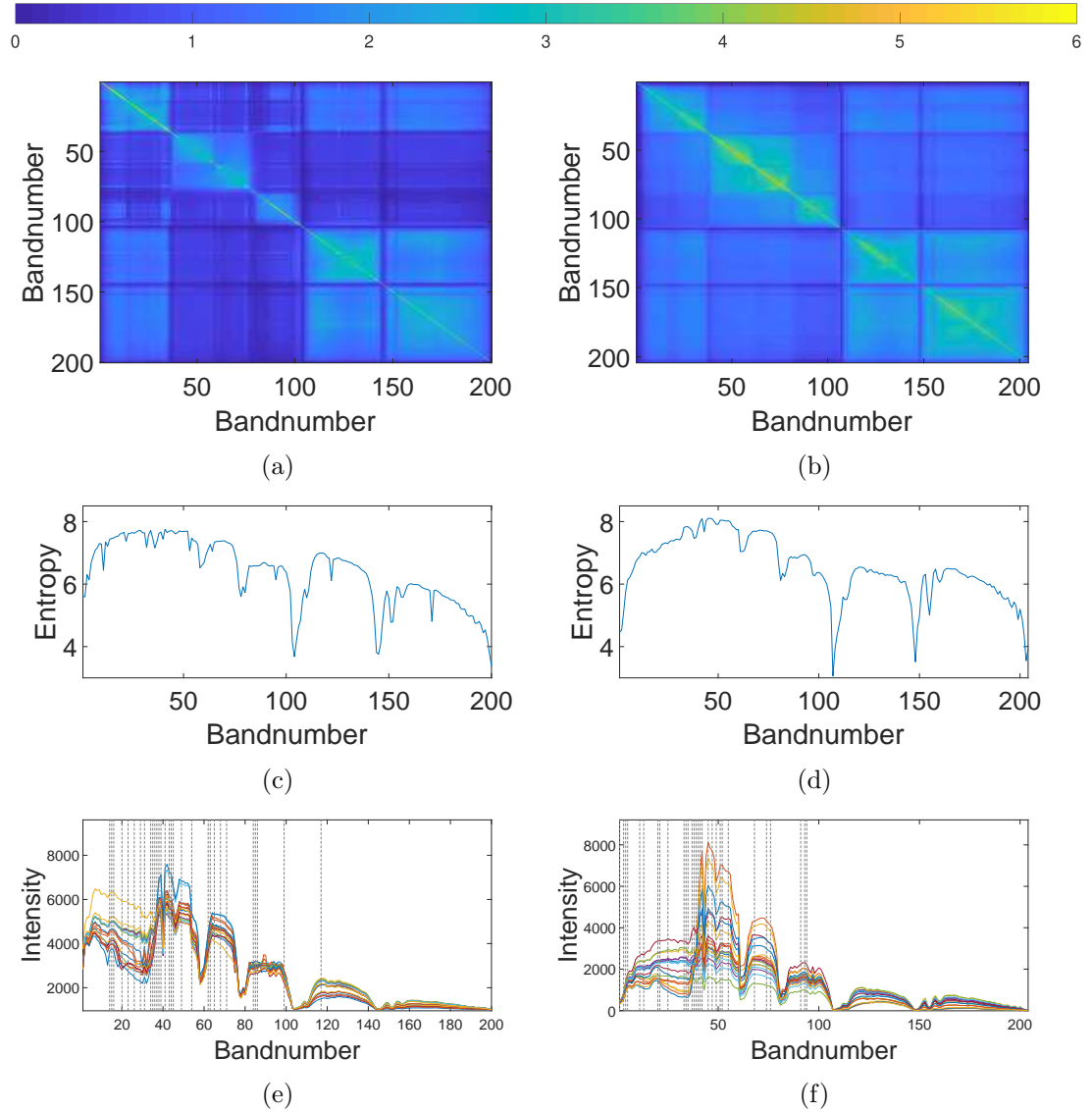
Figure 4.11: Comparison of mutual information matrix, entropy and resulting selected bands for the Indian Pines and Salinas datasets. The mean spectra of all 16 classes are indicated in different colours and the vertical lines indicate the selected bands.(a) - (b) MI, (c) - (d) entropy and (e) - (f) selected bands with DGSA where (a), (c) and (e) refer to the Indian Pines and (b), (d) and (f) to the Salinas dataset. The selected bands could achieve 83.7% OA for Indian Pines and 92.6 % for Salinas.
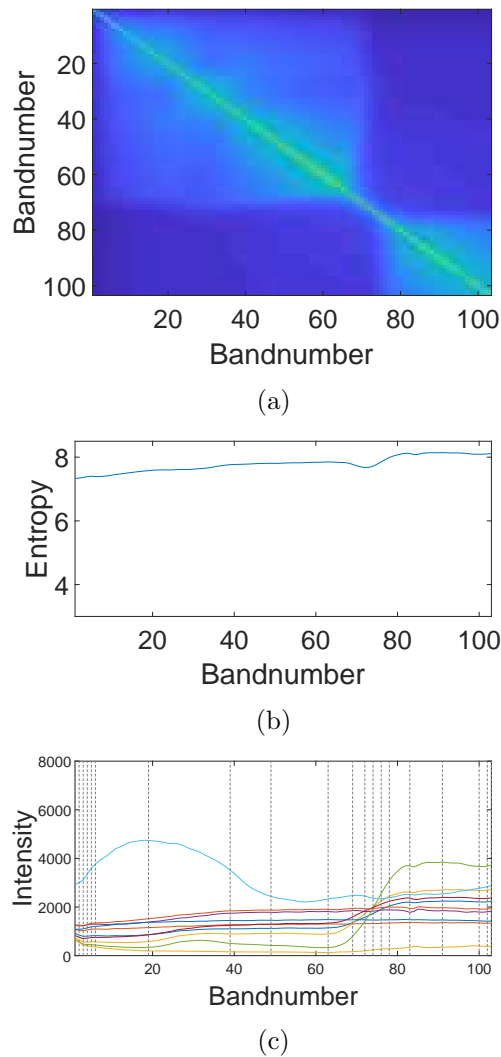
Figure 4.12: Comparison of mutual information matrix, entropy and resulting selected bands for the Pavia University dataset. The mean spectra of all 9 classes are indicated in different colours and the vertical lines indicate the selected bands. (a) MI, (b) entropy, and (c) selected bands with DGSA with an OA of 93.5%.

To verify the selected bands of MIMR-DGSA, the bands of a representative run were plotted on top of the mean spectra of each class in all datasets in Figure 4.11 and 4.12 alongside the MI matrices and entropy tables. For the Indian Pines dataset, bands 14 - 16, 20, 23, 26, 29, 31, 34 - 39, 41, 43 - 45, 49, 54, 62, 63, 65, 68, 71, 84 - 86, 99 and 117 were selected. In the Salinas dataset, bands 1, 3 - 5, 11, 13, 20, 21, 25, 33 - 35, 37 - 42, 45, 47, 49, 51, 52, 55, 68, 74, 76, 91, 93 and 94 were selected and for the Pavia

University dataset, only 20 bands were selected, i.e. 1 - 6, 19, 39, 49, 63, 69, 72, 74, 76, 78, 83, 91, 100, 102 and 103. The wavelengths from number #40 to #100 in both the Indian Pines and Salinas scene have the highest entropy and equally a relatively low MI with other regions. Looking at the spectra, this region seems to show the biggest differences between the classes, justifying why most bands are selected in this area. The bands edging the water absorption regions have a low MI but also a very low entropy, which is why they are not selected. The bands higher than 120 have a relatively high entropy but also a high MI with adjacent bands over a larger region. This is likely the reason why they are hardly ever selected. The mean spectra seem to be very similar in this region as well for most classes. The Pavia University dataset contains the most differences between the classes in the exact wavelength region around number 70, with a slightly lower entropy but also a much lower MI with all other regions. This makes the corresponding region most significant for classification. The edges of all spectra also show some changes in the shapes and are therefore selected. The rest of the bands seem to be evenly distributed over the spectral range.

4.2.4   Summary

In this section, the MIMR-DGSA algorithm was proposed for hyperspectral band selection. The algorithm is based on the MIMR criterion aiming to maximise the entropy of bands and minimise the MI between the bands in a subset. To evaluate the criterion, the entropy and MI are pre-calculated. A detailed analysis of the calculation of these measures for hyperspectral bands was presented resulting in the VarBWFastMI algorithm that calculates the pairwise MI of hyperspectral datasets in a reasonable amount of time and incorporates variable bandwidths for the density estimation. Additionally, a modified DGSA is developed introducing a neighbourhood and movement concept for feature subsets in a discrete optimisation space that generates potential subsets in a heuristic way and eventually generates a pseudo-optimal solution. Results on three publicly available HSI remote sensing datasets show that the proposed MIMR-DGSA performs very similar to the original MIMR-CSA, on which it is based, but is much faster and easier to use, as it only has three parameters. Parameters for the num-

ber of agents and iterations are directly correlated with the quality of the optimisation, whereas the number of candidates for the initialisation requires individual tuning based on the underlying dataset.

When comparing both CSA and DGSA with other EAs, it could be shown that FA and PSO are both less suitable to optimise the MIMR criterion. The effect of this is that the achieved classification accuracy is subject to high variance. While PSO and FA can achieve a higher accuracy in individual cases, DGSA was shown to perform superior and be more robust in average over all datasets. This also lead to the conclusion that the MIMR criterion itself does not necessarily optimise classification accuracy. Furthermore, PSO and FA require almost as many parameters as CSA and need therefore additional tuning. None of the examined algorithms performs consistently best in terms of accuracy, however, DGSA poses a robust, faster and easier to use alternative to CSA, which both optimise the MIMR criterion best. In combination with VarBWFastMI, an efficient and robust state-of-the-art band selection algorithm was proposed.

4.3   Embedded band selection using sparse autoencoders

In the previous section, the focus was on unsupervised band selection using a filter method. This section will introduce a novel embedded method that works equally unsupervised but the band selection is implicitly realised in the training process of a learning algorithm based on autoencoders. Autoencoders are neural networks that reproduce the input at the output by learning an abstract representation of the data. Neural networks work with weights on the connections between neurons. The relevance of individual neurons can be implicitly identified looking at the magnitude of those weights. By putting a sparsity constraint on the input weights, it is possible to identify those input features that contribute most to the reconstruction, which can in turn be utilised for band selection when reconstructing hyperspectral data. Additionally, it is proposed to segment the input data into several spectral regions. This enables the training of several, parallel and independent autoencoders with the advantage of a

faster convergence and a band selection that covers all spectral regions to appropriate measures. To determine the contribution of the spectral regions, the DD of each region is calculated and the desired number of bands distributed accordingly. This achieves a good distribution over all spectral regions whilst taking the individual weight of the regions into account.

4.3.1 Proposed approach

Basic principle

The proposed Segmented Autoencoder Band Selection (S-AEBS) algorithm consists of several steps. At first, the hyperspectral data is analysed and segmented into several logical spectral regions. For each segment, the DD is calculated indicating percentage of bands required for that segment. The number of desired bands can be distributed accordingly among these segments. For each segment, an AE is trained and the corresponding number of input bands with the highest weights are selected. The results of all AEs are then concatenated to form the ultimate set of selected bands. The procedure is visualised in Figure 4.13.
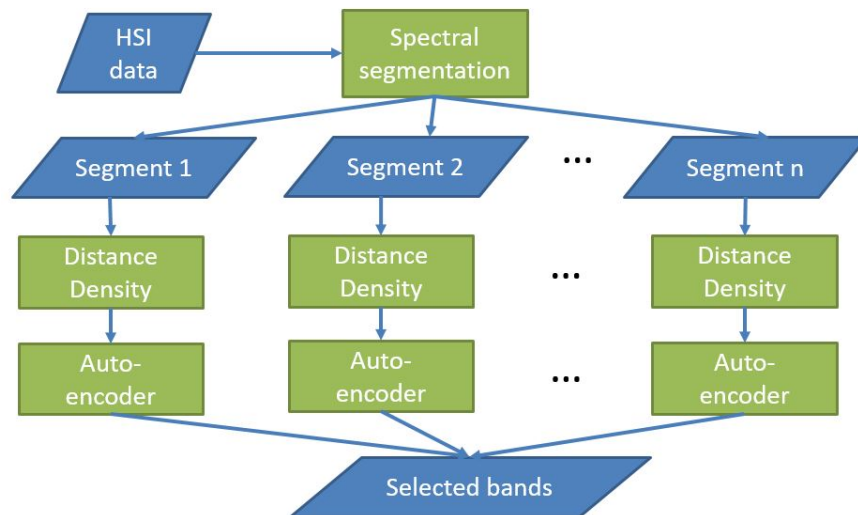
Figure 4.13: Procedure of S-AEBS algorithm. The input data is split into several logical segments which are then treated separately. The results are ultimately concatenated.
.

The AE proposed in [62] works by the following principle for feature selection. By minimising the reconstruction error of the output of the AE, the weights $\mathbf{W}^i$ that are used to connect the neurons are adjusted in each iteration. With convergence of the AE, the weights of the input layer $\mathbf{W}^1$ implicitly give information about the contribution of each individual input feature. This information can be used to identify the most important input features by ranking them according to their weight. To enforce this effect, [62] suggest to add a row-sparse regularisation term on the input weight matrix $\mathbf{W}^1$ which is realised by the $L_{2,1}$ norm:

$$\|\mathbf{W}^1\|_{2,1} = \sum_{i=1}^{d} \sqrt{\sum_{j=1}^{h} |w_{ij}^1|^2} \tag{4.10}$$

The i-th row $\mathbf{w}_i^T$ of $\mathbf{W}^1$ corresponds to the weights of all connections from the i-th feature and $|\mathbf{w}_i|$ gives indication on the contribution of the i-th feature to the reconstruction. The resulting loss function is defined as

$$\mathcal{J}(\theta) = \frac{1}{2m}\|\mathbf{X} - \hat{\mathbf{X}}\|_2 + \alpha\|\mathbf{W}^1\|_{2,1} + \frac{\beta}{2}\sum_{i=1}^{2}\|\mathbf{W}^i\|_2 \tag{4.11}$$

where α is a trade-off parameter between the reconstruction loss and the sparsity regularisation. An additional weight decay term is added with β being the penalty parameter, and m denotes the number of samples available for input matrix $\mathbf{X}$. This term prevents overfitting and enforces convergence of the optimisation. After optimisation, the bands are indicated by the norms of the columns of the input weight matrix $\mathbf{W}^1 = (\mathbf{w}_1\mathbf{w}_2...\mathbf{w}_d)$ for d input bands, where $\arg\max_i |\mathbf{w}_i|$ indicates the feature with the highest relevance.

In a hyperspectral dataset, each pixel's spectrum can be used as an input to the AE. The selected features from the defined AE represent those bands that are most relevant to the reconstruction of the spectrum and can be interpreted as the most informative bands. The functionality is depicted in Figure 4.14.
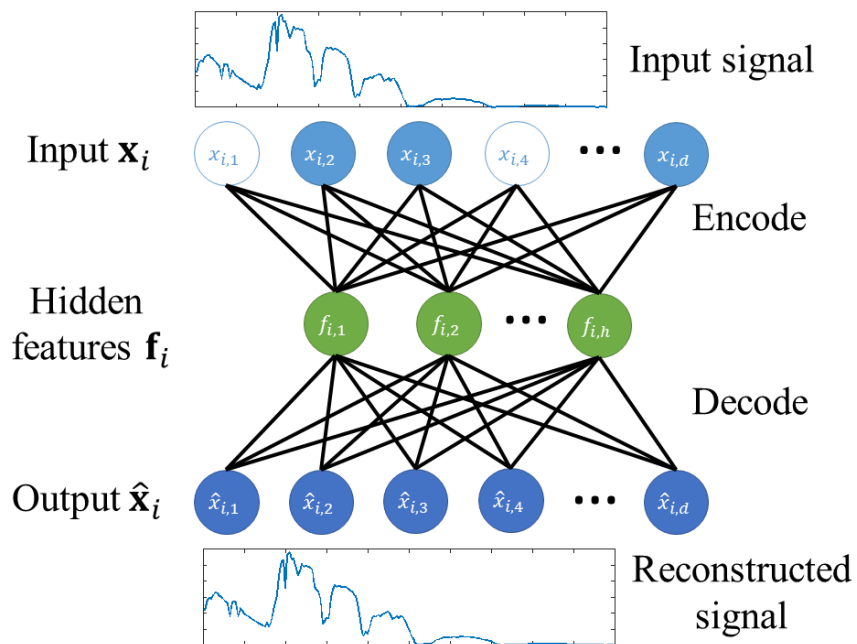
Figure 4.14: Schematic of AE band selection. The AE is trained to reconstruct the input spectra. With progression of the training process, the $L_{2,1}$ norm enforces sparsity on the weights of the input bands. Bands with higher weights (indicated by the filled nodes in the input layer) are considered as contributing more to the reconstruction and are therefore selected by the algorithm.
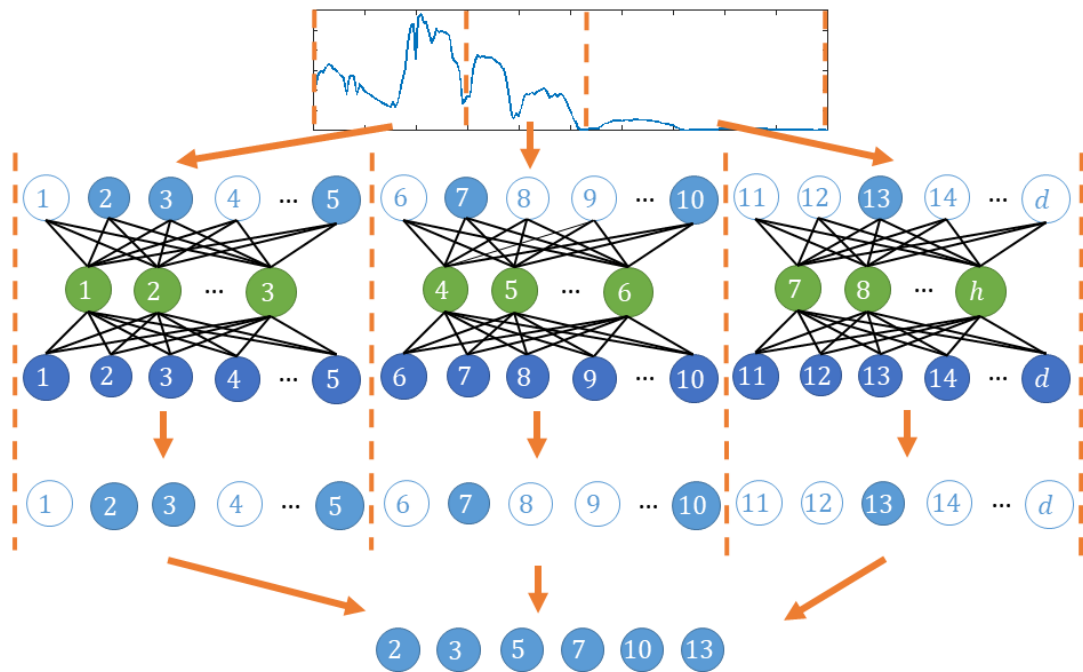
Figure 4.15: Schematic of S-AEBS. The filled nodes in the input layer indicate the input bands with the highest weights and are selected by the algorithm. For each spectral segment, a separate AE is trained, the bands with the highest weights in the input layer selected, and the results are concatenated to form the final selected band subset.

Segmentation of spectral regions

The spectral region covered by the utilised sensor can usually be divided into several logical segments and each of these regions contain a different amount of information about the dataset. Other algorithms, such as Segmented PCA [38] have adopted this concept successfully in the past. These segments are commonly generated by looking at the correlation matrix of the spectral bands. More information about the segmentation will be given in Section 4.3.2 on the example of specific datasets. Once the dataset is segmented into spectral regions, one AE for each segment can be trained and the resulting bands of each segment are simply concatenated. This process is visualised in Figure 4.15

As mentioned above, each spectral segment likely has a different amount of information necessary for the reconstruction. To account for this, the concept of DD from [59] is adopted here. The DD dd_i for segment i with m samples and n bands is defined as:

$$dd_i = \frac{1}{n-1}\sum_{j=1}^{n-1}d_j; \ d_j = \sum_{k=1}^{m}|r_{j+1k} - r_{jk}| \tag{4.12}$$

where d_j is the absolute difference between the reflectance values r_{jk} of two adjacent bands j and $j+1$ in sample k.

The number of bands n_{b_i} for the i-th segment can then be calculated by:

$$n_{b_i} = \frac{dd_i}{\sum_{i=1}^{s}dd_i} \times n_b \tag{4.13}$$

where s is the number of segments and n_b the total number of desired bands.

According to this calculation, spectral segments with a high DD contribute more selected bands to the final subset than segments with a low density. The assumption made is that spectral regions, where adjacent bands show a high variance in intensity indicate a higher information content.

4.3.2 Experimental results

The algorithm was tested on the three benchmarking HSI datasets Indian Pines, Pavia University and Salinas as specified in Appendices A.1, A.3 and A.2 respectively.

Segmentation

Choosing the right segments has significant impact on the classification performance [8]. Other than in [59] where the overall spectral region is divided into several equally sized segments, logical regions dependent on the specific dataset are attempted to be identified. For the purpose of segmentation, the Indian Pines and the Salinas datasets are considered to be the same, as they are acquired by the same sensor and depict similar objects. The correlation matrix can help with the segmentation. It is depicted for the Indian Pines dataset in Figure 4.16 and for the Pavia University dataset in

Table 4.9: Number of bands selected for each segment of the Indian Pines dataset

# Bands	0 - 35	36 - 103	104 - 144	145 - 200
9	3	5	1	0
30	10	16	3	1
100	34	54	9	3

Figure 4.17. Alongside, the mean spectra of all classes are shown to further verify the choice of regions. For the Indian Pines data, the segmentation points lie at 725 nm, 1400 nm and 1940 nm. The first segment from 400 - 725 nm is roughly the VIS spectrum, the second segment from 725 - 1400 nm is roughly the NIR region, the third from 1400 - 1940 nm roughly the SWIR and the last everything above. These spectral segments do not precisely adhere to the definitions of the above mentioned regions, but they indicate that the segmentation specified from visual inspection of the correlation matrix is justified as logical segments by their physical interpretation. For the Pavia University dataset, the segmentation point is roughly at 730 nm, which lies within the so-called *red edge*, where a significant change of reflectance in chlorophyll can be detected. This likely explains the change in spectra in this area and therefore also justifies the segmentation, because image regions with chlorophyll will have a very different spectral reflectance in both regions.

Based on the DD from Equation 4.13, the number of bands for each segment depending on the total number of desired bands can be calculated. Examples for the distribution between the segments for different numbers of bands for the Indian Pines dataset can be seen in Table 4.9. One can see that the first two segments contain significantly more information than the last two. Judging from the spectral signatures in Figure 4.16c, the first two segments display much more variance between the classes and between adjacent bands than the other segments.

Performance analysis

To assess the performance of the proposed algorithm, it was tested in various configurations. As a means to evaluate the influence of the segmentation, Autoencoder Band
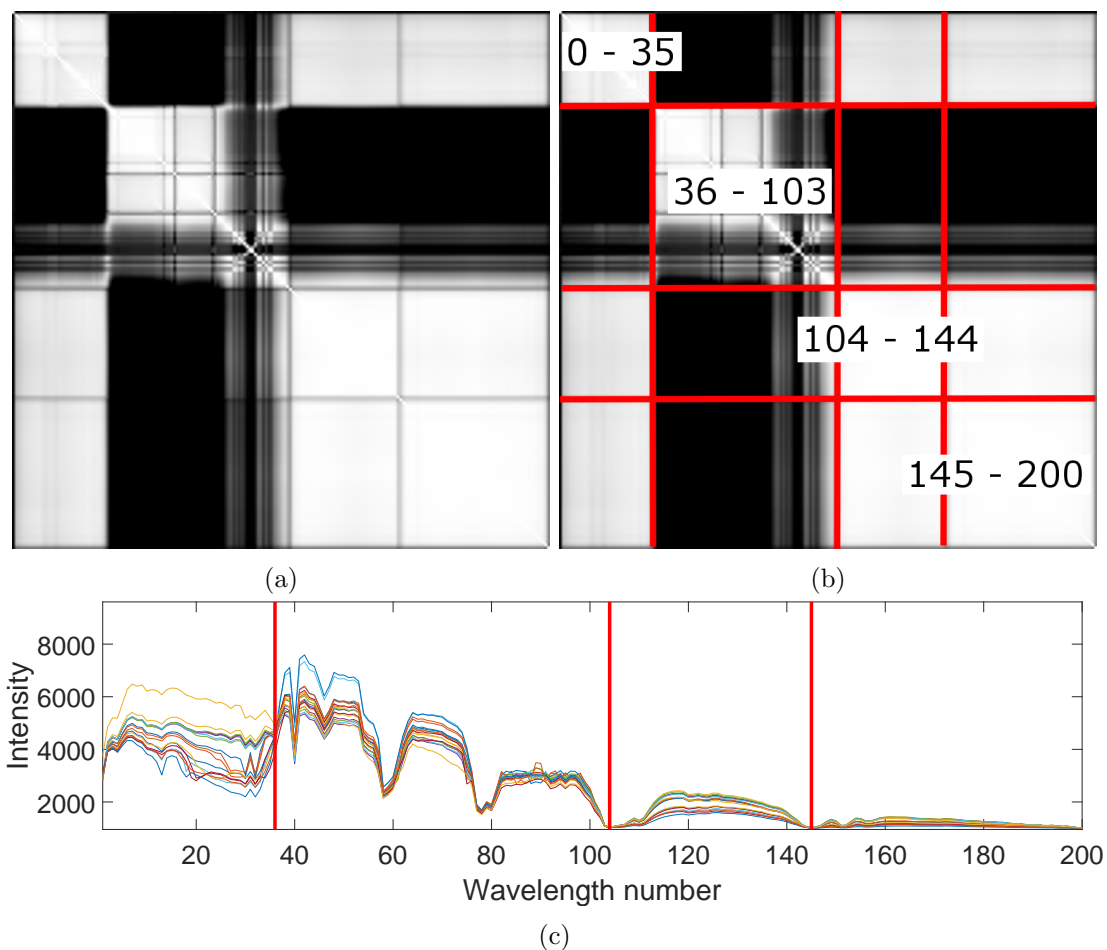
Figure 4.16: Visualisation of the spectral segmentation of the Indian Pines dataset. (a) Correlation matrix, (b) selected segments, and (c) the mean spectra per class with according segments.
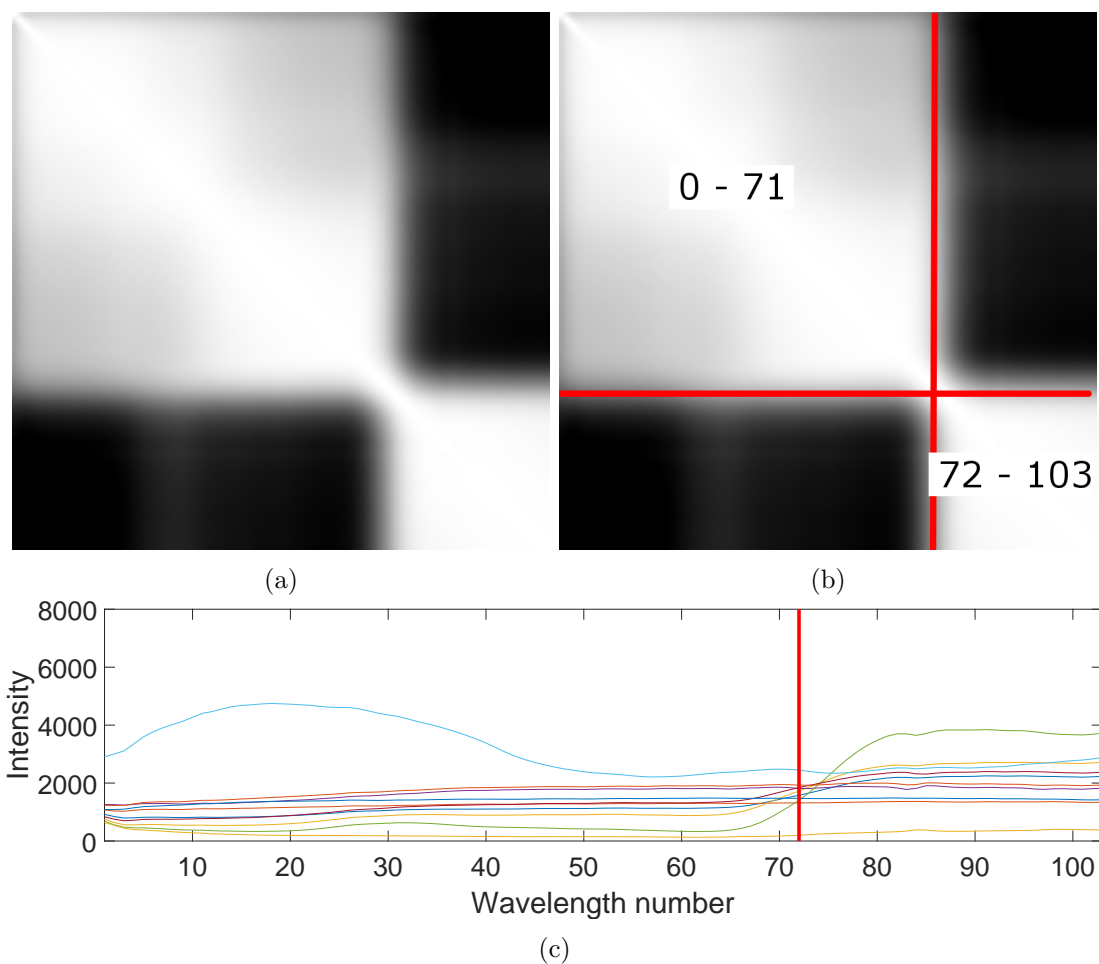
Figure 4.17: Visualisation of the spectral segmentation of the Pavia University dataset. (a) correlation matrix, (b) selected segments, and (c) the mean spectra per class with according segments.

Selection (AEBS) without segmentation was performed and compared with the results of S-AEBS. The parameters α and β of S-AEBS were tuned between 10^{-2} and 10^{-5} for each dataset individually. The hidden layer has the same number of nodes as the input layer, every Segmented Autoencoder (S-AE) is terminated after 10,000 iterations and the AE without segmentation takes longer to converge and is therefore terminated after 30,000 iterations. These values were determined empirically. ϕ^1 is set to a sigmoid function σ and ϕ^2 is set as linear activation, according to [62]. As an additional test to verify the effects of segmentation, random features of each segment were selected based on the segmentation and distance density, labelled S-RandBS. To benchmark the overall performance, the algorithm was compared with WaLuMI [136] and MIMR-CSA from Section 4.2. MIMR-CSA was applied using the same specifications as in Section 4.2.

Computational complexity and runtime measurements

The fact that MIMR-CSA has a computational complexity of $\mathcal{O}(i \times n_{cl} \times n_b^2)$ was established in Section 4.2.3, which is linearly dependent on the number of iterations and the population size but quadratically dependent of the number of desired features because of the MIMR evaluations. According to [122], WaLuMI has a computational complexity of $\mathcal{O}(l^2 \times n) + \mathcal{O}(l^3)$, where l is the number of bands available and n the number of samples in the dataset. The cubic runtime with respect to l stems from the clustering step in WaLuMI. The quadratic and linear runtimes with respect to l and n respectively are caused by the MI calculation for the dissimilarity space. The implementation used for WaLuMI uses the histogram density estimation technique, which is why the time consumption for the MI is still manageable. To analyse the complexity of AEBS and S-AEBS, the complexity of the learning phase of a feed forward neural network has to be considered. Determining the complexity of neural networks is still subject to research [139] and not a straight forward task and largely dependent on the design and architecture of the network. For a trained network, it is argued in [8] that the main dependency lies in the number of neurons and therefore neural connection between them, which increases exponentially with the number of neurons. Because

segmenting leads to AEs with less input neurons, their complexity also decreases because the unsegmented AE includes a large number of unnecessary neural connections. In the proposed approach, the complexity is additionally dependent on the number of iterations per training, as the band selection is realised during the training step. It is found that smaller autoencoders converge faster and therefore need less iterations for training. To summarise, MIMR-CSA is the only algorithm of the compared ones that is dependent on the number of desired bands, whereas all other algorithms depend on the number of desired bands. For S-AEBS, this dependency can be decreased compared to AEBS with the segmentation.

To measure the runtime, all algorithms were run on an Intel core i5 @ 3.2 GHz with 16GB Memory. S-AEBS and AEBS are both implemented in the TensorFlow framework for Python, MIMR-CSA is implemented in Matlab and for WaLuMI, the C++ implementation provided in [136] is used. Because of the different implementations and runtime environments, the times cannot be directly compared. However, general trends and structural differences of the algorithms can be identified. Runtime measurements for the Indian Pines and Pavia University datasets are shown in Tables 4.10 and 4.11 respectively.

Table 4.10: Time measurements of all algorithms with different number of selected bands for the Indian Pines dataset in seconds

Bands	WaLuMI	MIMR-CSA	AEBS	S-AEBS
10	21.5	3.0	404.5	145.0
50	20.7	7.9	404.5	145.0
100	18.4	21.1	404.5	145.0
150	16.9	47.3	404.5	145.0

One can see that AEBS and S-AEBS have a significantly higher runtime than both WaLuMI and MIMR-CSA. This is largely due to the fact that neural networks are inherently very computationally expensive. MIMR-CSA additionally benefits from the pre-calculation of the MI and entropy values, whereas WaLuMI benefits form an

Table 4.11: Time measurements of all algorithms with different number of selected bands for the Pavia University dataset in seconds

Bands	WaLuMI	MIMR-CSA	AEBS	S-AEBS
10	21.3	3.0	208.8	74.6
50	25.0	8.5	208.8	74.6
75	22.0	17.2	208.8	74.6

efficient C++ implementation. S-AEBS and AEBS can be significantly accelerated using a GPU implementation that can potentially speed up the training process of the AEs dramatically. It however also becomes evident that S-AEBS is notably faster than AEBS because of the lower complexity of the smaller AEs and the resulting faster convergence.

Classification performance

To evaluate the quality of the selected bands, the reduced datasets were classified using an SVM with a Gaussian RBF kernel. 10% of pixels of each class were randomly selected for training, and the rest to test the classifier. Since the AE optimisation is done with random initialisation, each training process will produce a slightly different band subset. To account for this, 20 AEs were trained in both AEBS and S-AEBS and each of these subsequently classified with 5 SVMs with different training and testing samples, resulting in 100 runs. Class-wise accuracies for the three datasets can be seen in Tables 4.12, 4.13 and 4.14, where 30 bands were selected for the Indian Pines and Salinas, and 20 bands for the Pavia University dataset. Due to very different number of samples available for each class, accuracies may vary strongly and certain classes show a very high standard deviation. Overall, S-AEBS performs best in most cases, where the margin between S-RandBS and AEBS is very small for the Salinas and Pavia University dataset.

In Figures 4.18, 4.19 and 4.20, the OAs of the different algorithms with respect to the number of bands selected were compared. While there are notable differences between the three datasets, it becomes clear that WaLuMI performs worst in most

Table 4.12: Class-wise accuracies for individual algorithms on the Indian Pines dataset selecting 30 bands

Class	WaLuMI	MIMR-CSA	S-RandBS	AEBS	S-AEBS
1	60.5±15.8	58.0±12.7	54.0±14.9	**68.9±11.2**	64.8±14.0
2	**77.2±2.1**	72.6±3.1	73.4±4.2	61.3±5.0	74.8±2.5
3	63.3±1.7	64.3±3.3	60.2±6.5	40.8±3.9	**64.4±4.1**
4	65.2±7.1	59.5±8.2	58.3±7.2	52.8±7.3	**67.8±7.7**
5	87.7±4.3	**90.9±2.1**	89.5±3.0	86.7±4.5	86.9±3.7
6	93.3±1.2	**94.7±2.0**	93.6±2.0	94.4±1.3	92.3±2.1
7	**75.2±13.1**	73.4±14.1	64.4±20.3	55.8±22.4	72.9±18.8
8	95.6±2.5	97.0±1.6	96.8±1.6	**97.3±1.3**	96.8±1.7
9	**45.6±9.9**	42.9±15.2	37.1±13.8	26.4±18.2	30.9±16.3
10	**75.4±1.5**	70.6±4.9	70.8±5.4	54.3±9.0	74.3±4.5
11	80.5±1.8	81.4±1.6	80.1±2.2	77.9±2.9	**82.1±2.1**
12	70.7±2.2	70.6±3.0	67.8±5.8	38.7±6.5	**75.3±3.5**
13	93.6±3.9	95.7±2.8	94.6±3.5	**96.2±3.5**	95.5±2.7
14	95.0±1.9	95.3±0.9	94.7±1.7	95.5±2.1	**95.6±1.3**
15	48.0±2.2	50.0±6.0	**50.8±6.7**	44.0±4.1	49.5±5.6
16	84.1±4.5	82.7±6.4	**86.0±5.6**	81.3±6.6	84.4±5.8
OA	79.9±0.3	79.4±0.8	78.5±2.1	71.1±1.7	**80.4±0.8**
AA	**75.7±1.1**	75.0±1.8	73.3±3.1	67.0±3.0	75.5±1.9
Kappa	77.0±0.3	76.4±1.0	75.3±2.5	66.7±2.0	**77.5±0.9**

cases. For the Indian Pines dataset, S-AEBS outperforms the other algorithms for 30 - 80 selected bands and one can clearly see that the segmentation adds additional value to the quality of the selected bands. This can be seen on all three datasets. S-RandBS performs significantly worse on the Indian Pines dataset. However, on the other two datasets, S-RandBS performs almost identical to S-AEBS. For the Pavia University dataset, this is likely the case because most bands contain very similar information and choosing different subsets makes hardly a difference in terms of classification accuracy.

Table 4.13: Class-wise accuracies for individual algorithms on the Salinas dataset selecting 30 bands

Class	WaLuMI	MIMR-CSA	S-RandBS	AEBS	S-AEBS
1	99.3±0.3	99.3±0.2	**99.4±0.3**	**99.4±0.3**	**99.4±0.3**
2	99.5±0.4	**99.8±0.1**	**99.8±0.1**	99.7±0.2	99.8±0.2
3	99.3±0.4	98.1±0.7	99.1±0.5	**99.4±0.4**	99.3±0.4
4	**99.5±0.1**	99.4±0.4	99.4±0.2	99.4±0.3	99.4±0.2
5	**98.9±0.4**	97.3±0.7	98.6±0.5	98.8±0.5	98.8±0.4
6	**99.8±0.1**	99.7±0.1	**99.8±0.1**	**99.8±0.1**	**99.8±0.1**
7	99.7±0.2	**99.7±0.1**	99.7±0.2	99.6±0.1	99.6±0.1
8	88.3±1.1	87.1±1.0	**89.0±0.9**	88.6±1.0	88.7±1.0
9	**99.9±0.1**	97.7±0.5	99.8±0.1	99.8±0.1	99.8±0.1
10	**95.8±0.5**	94.1±0.8	95.4±0.8	95.7±0.6	95.7±0.6
11	98.0±0.9	90.1±2.9	97.8±1.1	**98.6±0.8**	98.3±0.9
12	99.7±0.2	99.4±0.5	99.7±0.3	**99.8±0.2**	**99.8±0.2**
13	98.0±1.8	99.3±0.5	**98.9±0.9**	98.6±1.2	98.8±0.9
14	96.9±0.7	**97.8±0.8**	97.8±1.4	97.5±1.2	97.6±1.1
15	70.1±2.3	67.6±2.4	**73.5±2.9**	72.6±2.3	73.4±2.3
16	98.8±0.5	98.7±0.4	98.8±0.4	**98.9±0.3**	**98.9±0.3**
OA	92.9±0.2	91.7±0.3	93.5±0.4	93.3±0.2	**93.5±0.3**
AA	96.4±0.1	95.3±0.3	96.7±0.3	96.6±0.2	**96.7±0.2**
Kappa	92.1±0.2	90.8±0.4	92.7±0.4	92.6±0.3	**92.7±0.3**

This is why most of the algorithms yield an almost identical OA. On the Salinas dataset, the algorithms show more variance. Still, S-RandBS and S-AEBS yield almost identical results, while both outperform AEBS. What this shows, is the impact the segmentation has on the quality of the selected bands. It forces the algorithm to pick bands in every spectral region and counteracts a sole concentration on regions with high information content like MIMR-CSA does as shown in Section 4.2.3. This becomes especially evident in the Indian Pines and Salinas datasets, where the segmentation improves the

Table 4.14: Class-wise accuracies for individual algorithms on the Pavia University dataset selecting 20 bands

Class	WaLuMI	MIMR-CSA	S-RandBS	AEBS	S-AEBS
1	**92.7±0.4**	91.3±0.9	92.6±1.2	93.1±0.8	92.6±0.8
2	96.5±0.4	**97.4±0.3**	97.2±0.4	97.0±0.5	97.2±0.6
3	73.0±1.8	71.5±1.9	**75.2±4.9**	75.2±3.7	74.4±3.8
4	90.7±0.9	**93.3±0.8**	92.9±1.0	92.4±1.3	93.0±1.1
5	**99.9±0.3**	98.9±0.3	98.8±0.3	98.9±0.2	98.8±0.3
6	78.6±2.5	**87.9±0.8**	84.6±3.2	83.1±4.2	84.7±3.2
7	77.4±1.7	73.3±3.6	**80.6±2.6**	81.7±2.5	80.2±2.6
8	86.9±0.7	85.7±0.9	**87.6±1.6**	87.5±1.7	87.4±1.7
9	99.7±0.2	99.9±0.1	99.8±0.2	99.8±0.2	**99.8±0.1**
OA	90.9±0.4	92.1±0.2	92.4±0.8	92.2±0.9	**92.4±0.7**
AA	88.2±0.6	88.8±0.5	**89.9±1.1**	89.8±1.0	89.8±0.9
Kappa	87.9±0.6	89.5±0.3	**89.9±1.1**	89.6±1.3	89.8±0.9

performance of AEBS. The fact that S-RandBS performs almost similar to S-AEBS on the Salinas dataset highlights that S-AEBS still has room for improvement. A possible explanation is the inability of S-AEBS to account for interactions between selected bands, as it considers only the weight of individual bands and ranks them accordingly. It is likely that the magnitude of the weights is not a suitable measure for ranking and it does not penalise high correlation between bands.

4.3.3 Summary

In this section, an AE design for embedded hyperspectral band selection is proposed. By putting a sparsity constraint on the input weights, spectral bands that contribute most to the reconstruction can be identified. Combining this approach with a segmentation of the spectral region and training several AEs results in a faster and better band selection than a regular AE. This forms the basis for S-AEBS that can compete with state of the art algorithms. While the time consumption of training several AEs is relatively
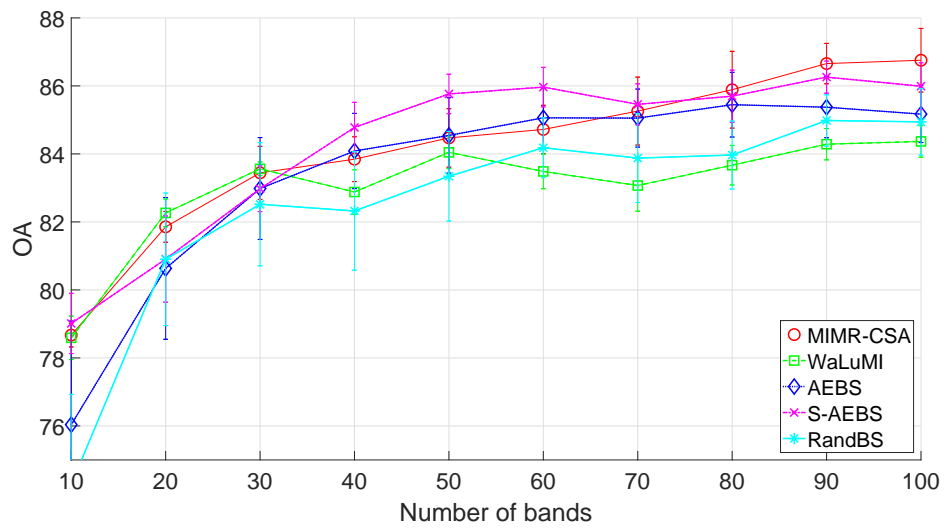
Figure 4.18: Classification accuracies for different algorithms on the Indian Pines dataset.
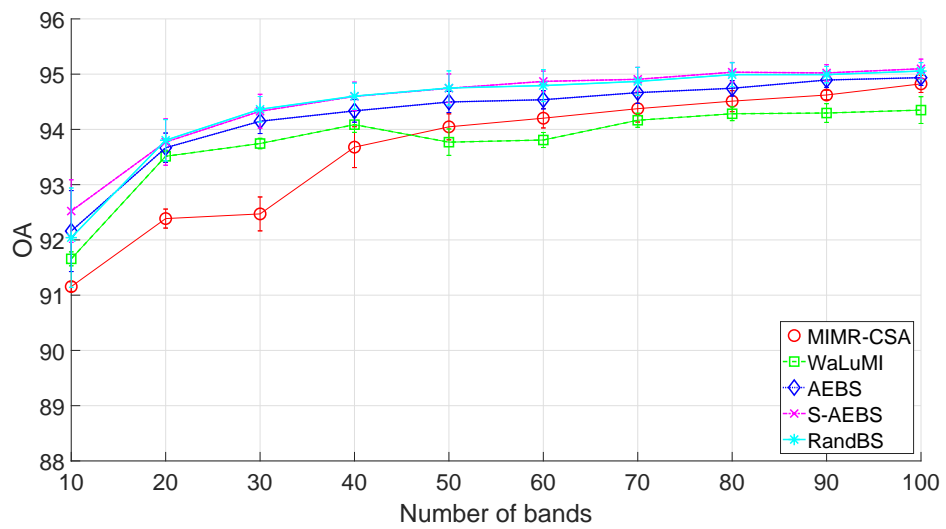
Figure 4.19: Classification accuracies for different algorithms on the Salinas dataset.
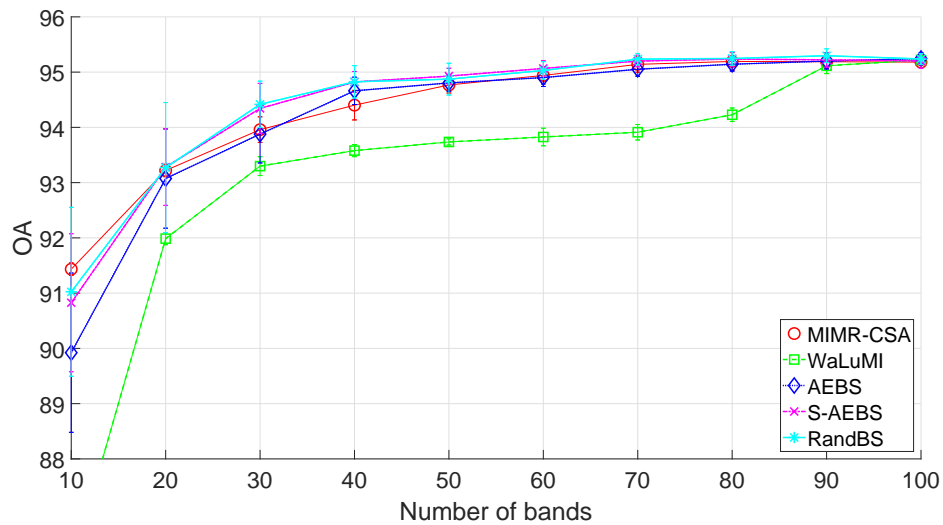
Figure 4.20: Classification accuracies for different algorithms on the Pavia University dataset.

high, Central Processing Unit (CPU) and GPU parallelisation of the S-AE training can be utilised to speed up the selection and improve the performance. Furthermore, comparable algorithms rely on the pre-calculation of information theoretic measures that can consume a considerable amount of time in itself. It was shown that S-AEBS benefits significantly from the segmentation step that forces the algorithm to select bands from all spectral regions. The performance in terms of classification accuracy on the selected bands is however not consistent over all datasets. In some cases, it performs equal to a random selection according to the underlying segmentation. This leads to the conclusion that S-AEBS is in the current state still subject to improvement. S-AEBS however outperforms MIMR-CSA and WaLuMI in most cases, which is why AE based band selection is potentially more efficient than the algorithm developed in Section 4.2.

4.4   Discussion

In this chapter, two band selection approaches were developed. The first one focused on an independent subset evaluation criterion based on information theoretic measures.

Information theory has in the past proven to be an effective measure to evaluate the information content of data without any prior information. The high computational burden of calculating these measures, specifically MI, poses a significant disadvantage. This problem was largely mitigated by pre-calculating values and developing VarB-WFastMI that drastically reduces repeated calculations. For algorithms that need to calculate measures like the MI on-the-fly during the runtime of the algorithm, this problem still prevails and highlights the limitations of these measures. Additionally, the optimisation capabilities of EAs with respect to an objective function were analysed. The efficiency of EAs is largely dependent on their internal structure, where better optimisation qualities usually come with a higher computational burden and a more complex definition. A relatively large number of hyperparameters additionally add to their complexity. Defining an algorithm tailored for hyperspectral band subset generation enabled the definition of DGSA that is characterised by a fast computation and low number of hyperparameters. What was also exposed is the fact that the analysed MIMR criterion is limited in its ability to produce optimal feature subsets. A superior optimisation can in some cases lead to a lower classification accuracy but on the other hand generates the most robust subsets over different datasets. The proposed MIMR-DGSA algorithm showed the highest robustness and flexible usage of all the analysed algorithms.

The second algorithm developed was based on the concept of embedded feature selection. AEs were analysed for their capabilities of hyperspectral band selection. Embedded band selection was of high interest as it shows high potential for very efficient algorithms. A single training instance of an AE can deliver detailed information about the input data and the relevance of individual bands. AEs, like most NNs however suffer from computationally very costly training periods. To mitigate this, the complexity of the AEs was minimised by segmenting the input data in to several spectral regions. As a side effect, this also increased the band selection qualities notably, as the algorithm is forced to select bands over the entire spectral region and not focus on a single region of disproportionately high information content. This became evident when comparing the proposed S-AEBS with the optimisation of MIMR. MIMR tends to focus strongly

94

on spectral regions with very high entropies, if the accompanying mutual information between neighbouring bands is appropriately low, and ignores regions with very low entropies. The segmentation in S-AEBS forces the algorithm to select bands in every region, even if it is a low number, which seems in turn to increase the quality of the subset notably. S-AEBS itself however has in the current state still room for improvement. Unlike MIMR, it is also much less transparent as to why certain bands are selected, whereas MIMR can be fairly straight forward interpreted. This is a common problem in NNs, where their superior performance is sometimes overshadowed by a less transparent solution.

Two very distinct algorithms were developed and analysed in this chapter. While S-AEBS outperforms other algorithms in classification accuracy, transparency and physical interpretability of the results are a clear advantages of MIMR-DGSA. The selected bands of both algorithms can be used to gain information about the chemical constitution of the imaged objects and as a result be used to develop sensing technology that focuses only on certain spectral ranges and thereby decrease the amount of data, the computational burden and most importantly the monetary cost of imaging systems. Although the algorithms showed that decreasing the number of bands can lead to a similar OA, the reduction is still not sufficient to build according MSI systems. Therefore, other approaches need to be considered to address the issue of constructing more cost- and data-efficient hardware. The next chapter elaborates on this concept and proposes a very cost effective solution to hyperspectral imaging.

Chapter 5

Neural Network Based Hyperspectral Reconstruction from LED Enabled Multi-channel Imaging

5.1   Introduction

Despite its maturity, one of the main reasons HSI technology is not yet widely applied in industry is its high cost. Depending on the desired spectral range, the price of an imaging system can vary from 10-20 thousand to over 100k GBP (as of March 2019). In addition to the pricing, the mode of operation often makes it not universally applicable. The most straight-forward mode for industrial applications is line scanning, which allows for relatively easy integration into production environments such as conveyor belts or even airborne systems where certain land areas are scanned linearly [22, 140–142]. This imaging method is however very prone to geometric distortions and therefore needs careful installation and control when implementing. HSI has a relatively low degree of automation and needs additional expert knowledge for data acquisition and analysis and thus very particular application specific parameter settings for the imaging

96

process [143] . A recent focus is put towards the development of compact cameras [144] that simplify the usage. One promising technique is snapshot imaging. Snapshot HSI is an increasing field of research, where techniques such as compressive sensing [23, 26, 27] allow for the integration of HSI datacubes within a single shot and therefore mitigate problems of geometric distortions. Coded aperture approaches utilise the principle of compressive sensing, which states that spectral scenes can be recovered from far fewer measurements than required by common line scanning or other sensors. These systems use a coded masked aperture to trade off spectral and spatial resolution for the advantage of sensing an entire spectral scene in a single aperture. Using computational imaging theory, the underlying data can be fully reconstructed. However, the resulting images typically only consist of a limited number of spectral bands, suffer from relatively low spatial and spectral resolution and therefore are often not suitable for the specific requirements of industry. The proposed system here is similar in the sense that it utilises computational imaging to reconstruct the data, but differs in the respect that it senses high resolution spatial images and requires very little and cost-effective optical elements for spectral sensing and reconstructs a full hyperspectral datacube with hundreds of wavelengths.

In the previous chapter, the advantages and limitations of band selection were dissected in detail. It was shown that a large majority of information in an HSI datacube can be captured by only a subset of all available bands. This information however is very application dependent and can vary strongly in different datasets, depending on the spatial absorption of the imaged objects. Rather than imaging a whole datacube and *a posteriori* removing unnecessary and redundant information, this chapter will focus on using application specific *a priori* knowledge to image only a subset of the data in the first place. MSI systems with variable settings for recorded wavelengths were developed in the past [145, 146]. In Chapter 4, it was shown that while a subset of the available bands can produce similar classification results to the original full set of bands, these subsets typically still require a large portion of the original data (50 or more bands) rendering it almost impossible for MSI camera systems to capture this amount of data. As a result, limiting oneself to a certain number of selected

wavelengths that can be captured by the respective MSI system might deteriorate the
predictive power of the system compared to the full hyperspectral information. On
the other hand, a small portion of the full spectral information might be sufficient
to accurately reconstruct the original spectra. The reconstruction of hyperspectral
data for various RGB and MSI systems has been a recent research focus that has
yielded some very promising results [147–149]. In this chapter, a multi-channel system
based on time-multiplexed RGB LED illumination is proposed. Regular MSI images
consist of several images at distinct wavelenths. The system proposed differs in this
respect. The RGB illumination used here only emits light at three wavelengths, which
are combined to different ratios to generate several colours, as detailed in Section 5.2.1.
The advantages of the system are low-cost, high mobility, high framerate, and improved
efficiency in hyperspectral data reconstruction. It is looked at in detail whether tunable
RGB LED illumination can provide additional information that is not captured by
regular RGB imaging technologies, when creating more channels. In mapping the
collected data to HSI prior by training a relatively shallow neural network, it is aimed
to realise a computationally efficient algorithm for HSI data reconstruction. Defining
a shallow network has the advantage of being computationally relatively inexpensive
and therefore realisable on embedded hardware.

5.2   Proposed system

The proposed system comprises of two main components. At first, multi-channel and
hyperspectral data of the same objects are acquired. This is used to train a reconstruc-
tion algorithm, that is then able to reconstruct hyperspectral data from unknown input
containing objects with the same spectral characteristics. This is visualized in Figure
5.1. The individual components are detailed in this section.

5.2.1   Hardware architecture

The architecture of the proposed system is schematically represented in Figure 5.2.
It comprises a Basler ace U monochromatic camera [150] with a spatial pixel resolu-
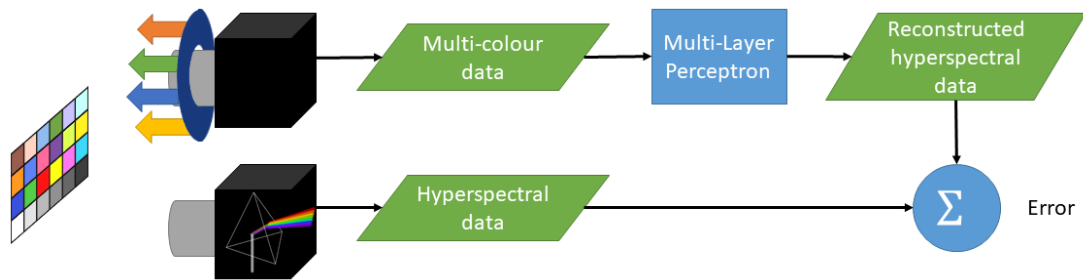
Figure 5.1: Procedure of the entire reconstruction setup.

tion of 1280 x 1024 pixel (HD), a framerate up to 200 fps and a USB 3.0 interface. The image integration is triggered by an Arduino Uno [151], which also controls an Adafruit Neopixel ring [152] containing 24 WS2812B based RGB LEDs. The LEDs are programmed to produce eleven different colours in sequence and the image acquisition is triggered by the switch of colour, realising time-multiplexed illumination. The image data from the camera is recorded by the Single Board Computer (SBC) Odroid XU4 [153]. Depending on the number of colour channels used, this allows for a recording speed from 18 (eleven channels) up to 200 (one channel) images per second, which allows the system to operate in video framerate. Notably, the components amount to a cost less than GBP 1000 (as of March 2019).

The selected eleven colours were chosen to emulate wavelengths between 400 - 650 nm matching the CIE 1931 standard observer. The colours were selected to be approximately equally spaced at 25 nm in terms of wavelength in that region. The settings for the intensities of the three colour channels are illustrated in Figure 5.3. Note that RGB LEDs can only generate colours by mixing intensities of the three RGB lights with very distinct peak wavelengths at 465 - 467 nm (B), 522 - 525 nm (G) and 620 - 625 nm (R) [20]. The irradiance of the generated colours was measured with a spectroscope and the results are shown in 5.3. It is apparent that the peak wavelengths remain the same, while the intensity combinations vary. A phenomenon called colour metamerism is exploited, stating that the human eye can perceive colours with different, non-matching spectral power distributions as the same colour [154]. Effectively, colours are generated that are perceived as the same colours as the desired wavelengths between 400 - 650 nm would appear as. It is believed that varying the illumination colour (or intensity
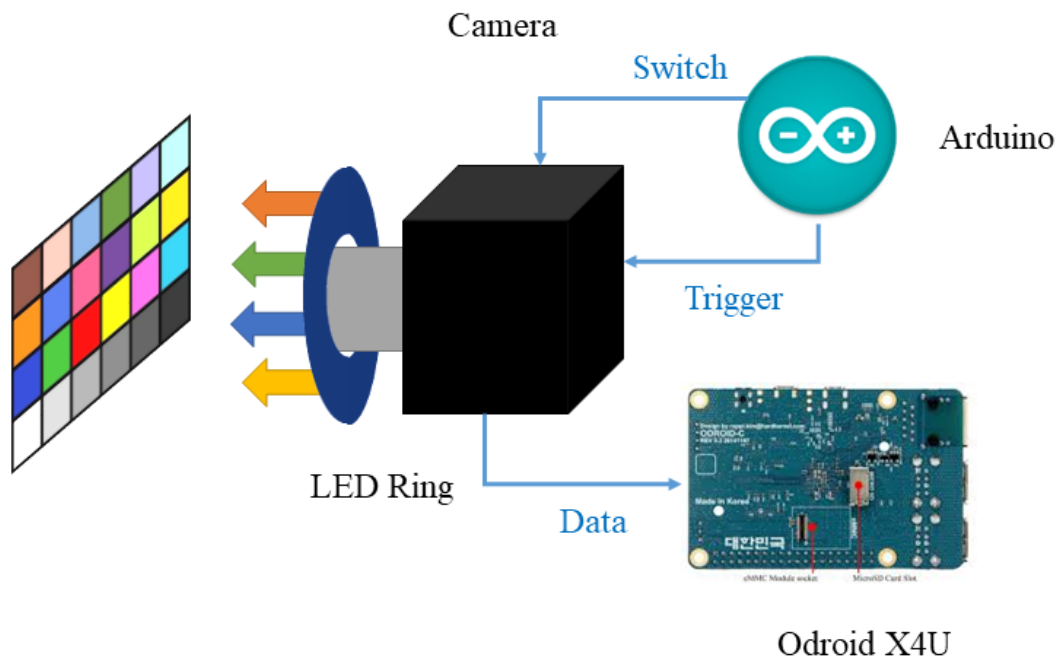
(a)
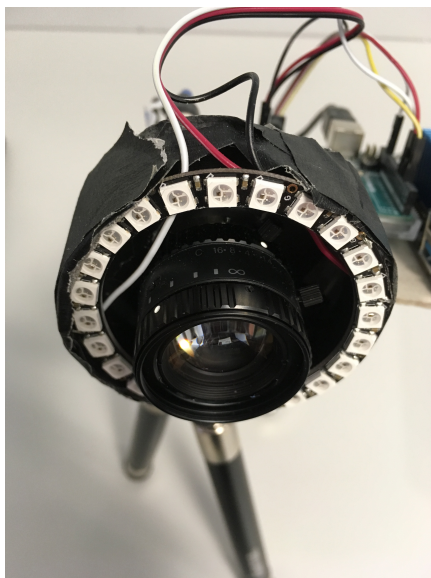
(b)
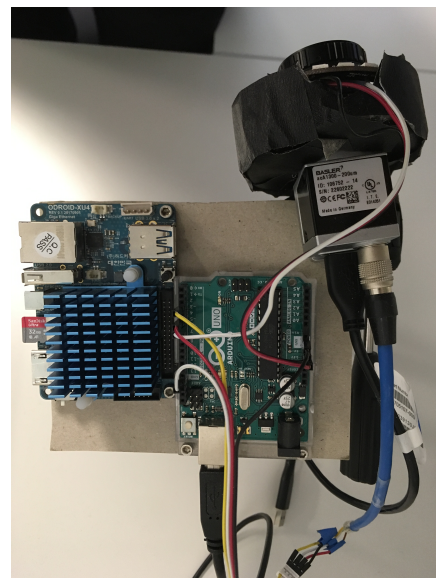
(c)

Figure 5.2: Architecture of the multi-channel system. (a) Schematic architecture, (b)
front view with lens and LED ring and (c) top view showing all components.
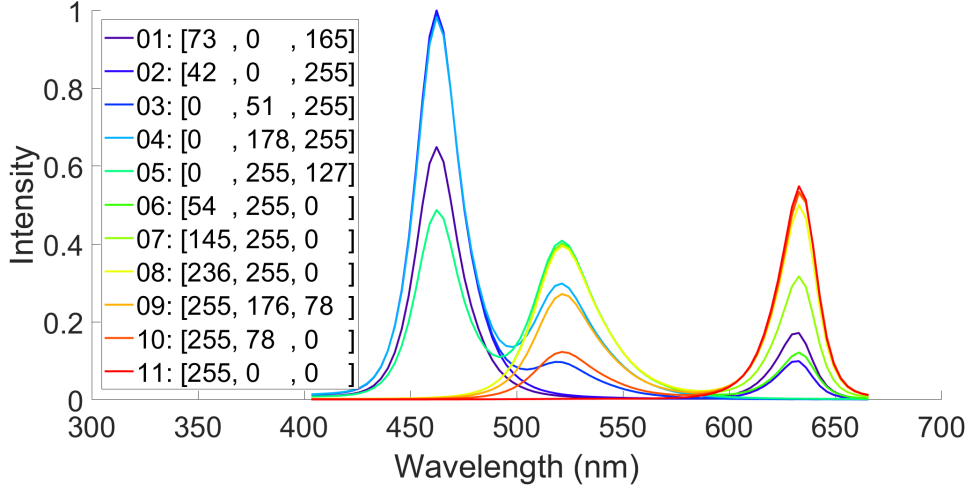
Figure 5.3: Measured irradiances of the eleven used colour configurations. The eleven configurations in terms of R, G and B values are indicated.

combinations), even without generating new wavelengths, increases variability within the data and therefore the accuracy of the predictive system.

5.2.2 Hyperspectral reconstruction

The reconstruction is performed by training an MLP defined as follows. An input matrix $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n\}^T$ of multi-channel pixel vectors $\mathbf{x}_i$ is mapped to an output matrix, $\mathbf{Y} = \{\mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_k\}^T$, where k is the number of objects and n is the number of total multi-channel observations. The MLP consists of four fully connected (FC) hidden layers, where the successor layer $l + 1$ to the l-th layer is calculated by

$$\mathbf{x}_i^{l+1} = f^l(\mathbf{x}_i^l) = \phi^i(\mathbf{W}^l\mathbf{x}_i^l + \mathbf{b}^l) \tag{5.1}$$

with activation function ϕ^l, weight matrix $\mathbf{W}^l$ and bias $\mathbf{b}^l$. The detailed architecture is visualised in Figure 5.4. The fact that the input layer has no activation function, i.e. $\phi^1 = f(\mathbf{x}) = \mathbf{x}$ enforces the use of all colour channels to further reduce fluctuations. The other activation functions are set to the sigmoid function σ. σ is chosen here as it is a very commonly used and produces very good results. Further optimisation might look into choosing different functions such as ReLU.
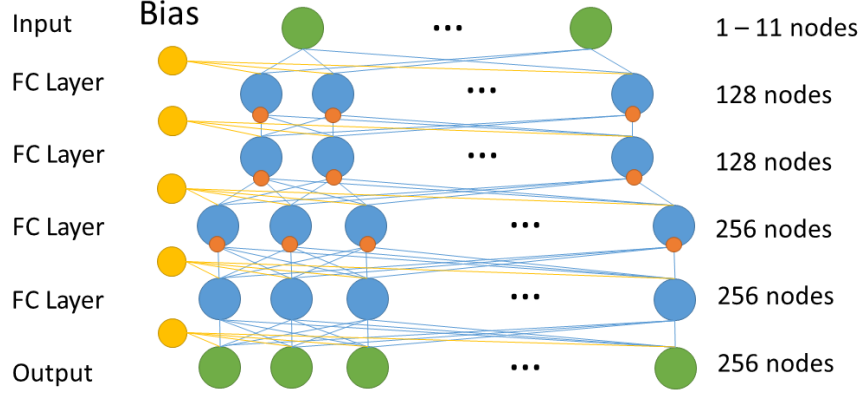
Figure 5.4: Design of the neural network. Orange dots symbolise sigmoid activation functions. The input matrix $\mathbf{X}$ can contain 1 - 11 nodes, depending on the number of channels recorded and is mapped to the output matrix $\mathbf{Y}$ through four fully connected hidden layers, where the first two consist of 128, and the last two of 256 nodes. Sigmoid activation functions are only applied on the first three hidden layers (indicated by the orange dots).

For training, a loss function $\mathcal{J}(\boldsymbol{\theta})$ of parameter vector $\boldsymbol{\theta} = \{\mathbf{W}^1, ..., \mathbf{W}^i, \mathbf{b}^1, ..., \mathbf{b}^i\}, i \in 1, ...5$ is defined. For the predicted output $\hat{\mathbf{Y}}$, a simple loss function can be defined as:

$$\mathcal{J}(\boldsymbol{\theta}) = \frac{1}{2n}\|\hat{\mathbf{Y}} - \mathbf{Y}\|_2 \tag{5.2}$$

The problem, however, is highly ill-posed, as the dimensionality of input vectors $\mathbf{x}_i$ is significantly lower than the dimensionality of output vectors $\mathbf{y}_i$ and slight changes in the input can translate into a severe alteration at the output. For this reason, weight decay regularisation is introduced into the loss function:

$$\mathcal{J}(\boldsymbol{\theta}) = \frac{1}{2n}\|\hat{\mathbf{Y}} - \mathbf{Y}\|_2 + \frac{\alpha}{2}\sum_{i=1}^{5}\|\mathbf{W}^i\|_2 \tag{5.3}$$

Furthermore, dropout is used while training the network for additional regularisation.

5.2.3 Implementation details

The implementation was done in the TensorFlow framework for Python, using the Adam Optimiser with the initial learning rate of 0.001. α was set to 10^{-6} and a batch size of 128 was used for training. A keep rate of 0.6 for the dropout was used and in

accordance with the definition of the popular AlexNet [120] only applied on the last hidden layer. Weight decay and dropout both minimise effects of overfitting and consequently enforce a relatively smooth output which satisfies smoothness characteristics of hyperspectral data. For optimally smooth results, it is empirically established that 5,000,000 iterations are necessary.

5.3 Experimental results

To verify the system, the proposed camera was used to image 20 differently coloured fabrics such as cotton, wool and polyester in various shades including four different blacks, four very bright materials in white, grey, cream and bright yellow, two blue shades, turquoise and two green shades and seven different red, orange and pink shades. These fabrics were chosen to test for the discriminability between subtle changes in colour. Additionally, the Macbeth ColorChecker chart with 24 colour patches and seven different organic objects, including red and white grapes, bananas, apples, pears, spinach leaves and tomatoes were imaged to include some naturally occurring materials. The imaged objects are illustrated in Figures 5.5, 5.6 and 5.7 for fabrics, organic materials and the Macbeth chart respectively. In total, 51 different objects were imaged with an HSI system covering the VNIR range from 400 - 950 nm in 256 bands. For each of the 51 objects, the mean HSI spectrum with 256 bands was calculated and used as the desired output of the MLP. The input was generated by imaging the same objects with the proposed multi-channel system. From these images, 4000 pixels were selected of each of the 51 objects and used as the input observations. All 4000 observations were mapped to the one HSI spectrum of the according class. The trained system is then able to recreate an HSI spectrum from any input, given that it is similar enough to the multi-channel inputs it is trained on. The multi-channel images were acquired in a dark room to minimise the effects of ambient lighting. For production use of the system, an ambient light image can be acquired and subtracted from all successive images to realise calibration. Both datasets were radiometrically calibrated and converted to reflectance using white and dark reference images, according to Section 2.2.3. In theory,

the multi-channel system only needs to be calibrated once, as the illumination is always constant. Given the influence of ambient light and the variable sensor sensitivity and LED radiance, it is still advisable to repeat the calibration before every imaging session.

Figure 5.5: RGB depiction of imaged fabrics.

In the first experiment, it was attempted to analyse the effect different numbers of bands have on the quality of reconstruction. The results for the EMD achieved after 1,000,000 iterations with 5 repetitions are shown in Figure 5.8. The bands were chosen to be as equally spaced as possible out of the eleven totally recorded bands. Interestingly, it can be seen that at least 6 colour channels are required to minimise the reconstruction error. This leads us to the presumption that adding more colour chan-

(a)

(b)

(c)
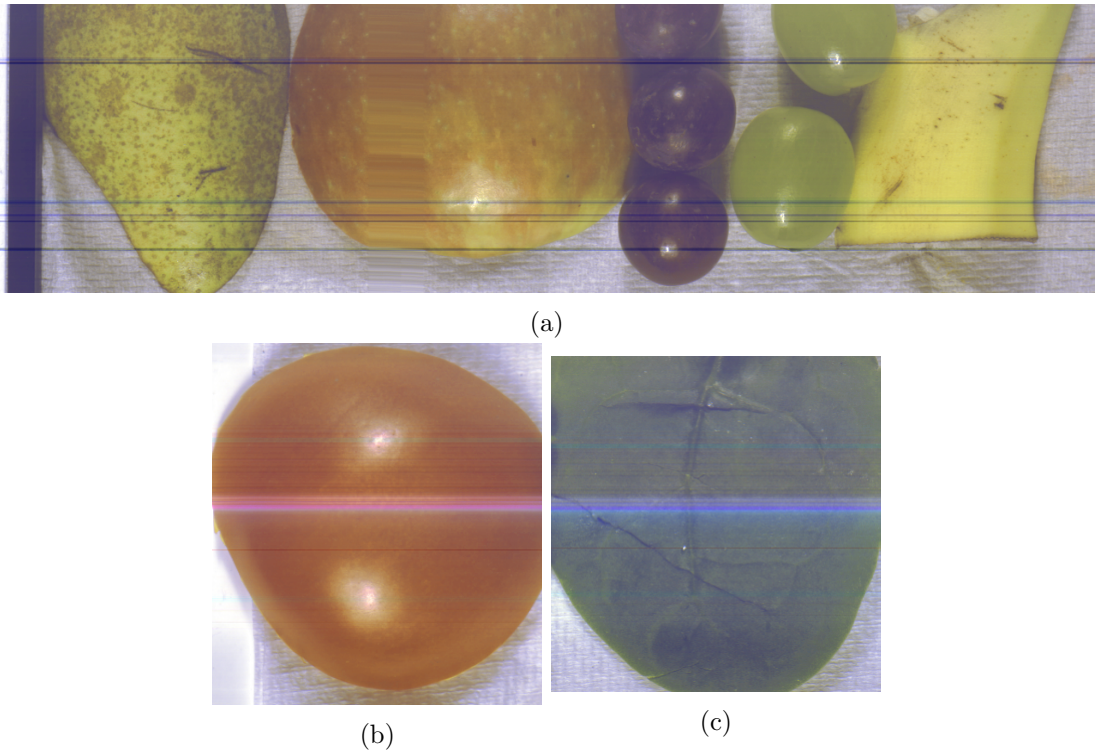
Figure 5.6: RGB depiction of imaged organic objects. From left to right: (a) Pear, apple, red grapes, green grapes and banana, (b) tomato, and (c) spinach leaf.

nels to RGB also adds valuable information to the reconstruction. According to these results, the reconstructed spectrum profiles from three, six and eleven bands respectively were compared in additional experiments. Both six and eleven channels produce an equally low EMD but recording less bands also potentially increases the framerate of imaging. Both were compared with three channel reconstruction to highlight the additional information gain over RGB. The three channels that are closest to RGB were selected to highlight the information gain over plain RGB images.

For quantitative evaluation, the 51 imaged objects were grouped in three groups; coloured fabrics, organic materials, and Macbeth chart patches. The reconstruction was evaluated using the three different measures EMD, SAM and SID detailed in Section 3.2.2. Reconstruction results after 5,000,000 iterations with three, six and eleven channels are given in Table 5.1. The reconstruction using eleven bands performs generally the best. The coloured fabrics produce a higher similarity to the original HSI data,

Figure 5.7: RGB depiction of the Macbeth ColorChecker chart.

whereas both the fruit and vegetable and the Macbeth chart generate very similar errors
for both six and eleven channels. This shows that the fabrics can be reconstructed very
accurately, even though some of them only vary slightly in the shade of the respective
colour. What can also be observed is that both six and eleven channels produce a sig-
nificantly lower reconstruction error than the three plain RGB channels. This confirms
the presumption that adding more colour channels reduces noise and possibly adds new
information to the MLP and therefore improves the reconstruction significantly.

The reconstruction is visualised in Figures 5.9, 5.10 and 5.11. Only representa-
tive classes of objects were chosen that display certain attributes of the reconstruction.
The coloured fabrics have relatively distinct spectral signatures, due to differences in
material and colouring, and thus can be quite accurately reconstructed. For black fab-
rics, the initially low signals lead to a less accurate reconstruction, as shown by a less
smooth curve. The linen and the polyester display almost exactly the same spectral
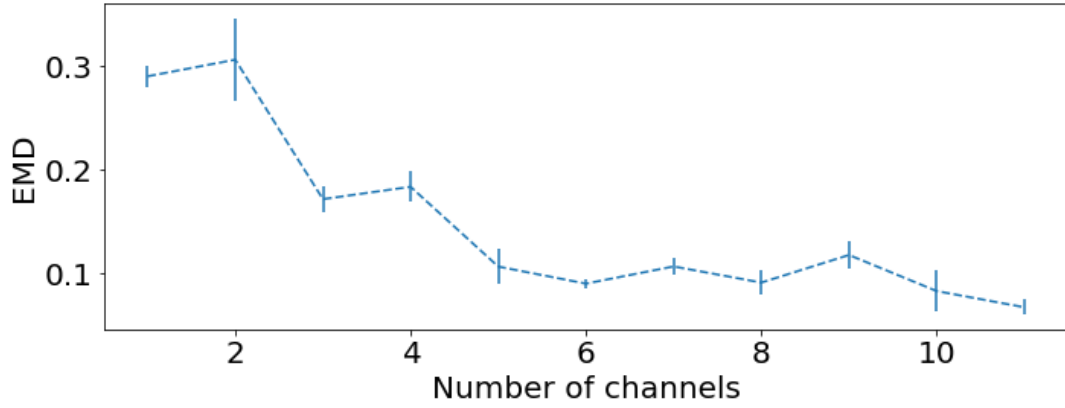
Figure 5.8: Reconstruction error for different numbers of channels.

Table 5.1: Reconstruction errors for the object groups. For 3 channels, channels 2, 6 and 11 were used according to Figure 5.3 and for 6 channels these are 1, 3, 5, 7, 9 and 11

	Fabric	Organic	Macbeth chart
	3 Channels		
EMD	0.475±0.500	0.947±0.630	1.279±0.279
SAM	0.077±0.028	0.114±0.053	0.122±0.011
SID	14.808±4.60	0.858±0.386	0.435±0.038
	6 Channels		
EMD	0.277±0.218	0.439±0.211	0.502±0.038
SAM	0.065±0.019	0.064±0.021	0.075±0.004
SID	15.60±3.420	3.594±1.317	2.412±0.405
	11 Channels		
EMD	0.128±0.095	0.228±0.143	0.219±0.032
SAM	0.032±0.015	0.034±0.013	0.046±0.003
SID	20.14±3.436	5.553±1.732	3.508±2.912

Table 5.2: Comparison of SVM classification using the original 256 bands with the reconstruction of these bands from three, six and eleven channels. For three channels, channels 2, 6 and 11 were used according to Figure 5.3 and for six channels, these were 1, 3, 5, 7, 9 and 11

	Original	3 Channels	6 Channels	11 Channels
Fabrics	98.80±2.89	86.84±4.88	91.89±3.28	93.10±2.92
Organic	99.81±0.63	91.81±1.78	98.62±0.65	98.86±0.57
Macbeth chart	98.49±2.89	98.55±0.63	99.92±0.08	99.91±0.14
OA	98.79±2.58	93.22±2.45	96.59±1.41	97.45±1.29

characteristics in the spectral coverage of the illumination, despite their differences in the spectral range from 700 nm upwards. The proposed camera system is likely not to pick up on these features as the illumination does not cover this spectral range. Consequently, the reconstructed spectra of these two classes of objects are almost identical. The metrical differences to the original spectra are however still small, which leads to the high similarity in Table 5.1.

For the fruit and vegetable, the input multi-channel data is more varied as the surfaces of the scanned materials are not homogeneous. As a result, the reconstruction shows some deviation from the original spectra but still adheres well to the HSI data. Specific features, such as the typical spike in reflectance from 680 - 730 nm caused by chlorophyll and referred to as the red-edge, is retained very well for the spinach.

For the Macbeth chart, the colours show very distinct spectral reflectance and can therefore be reconstructed very accurately, reflecting the nature and the usage of the chart. Difficulties arise for the grey (neutral) and black patches, where the low intensity of the signal and their flatness leads to errors in the reconstruction.

To further validate the quality of reconstruction, an SVM with Gaussian RBF kernel was trained to classify all 51 classes using 600 samples of each object with 5% of each class for training and the rest for validation. The parameters C and γ of the SVM were trained using a grid search with 5-fold cross validation. The 51 classes were again grouped and the mean pixel-wise OA was calculated for each group and summarised in Table 5.2.
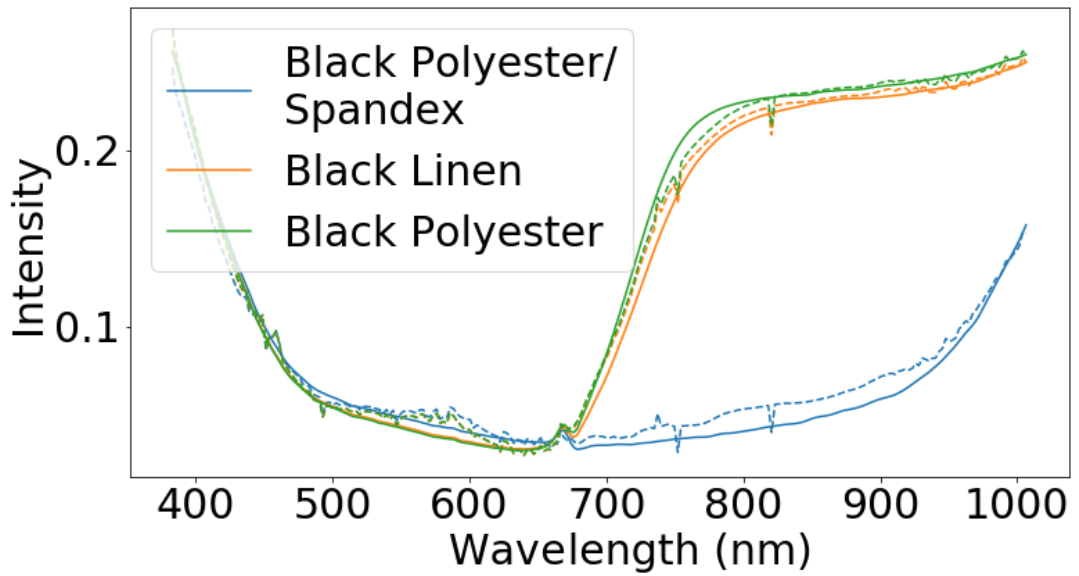
(a)
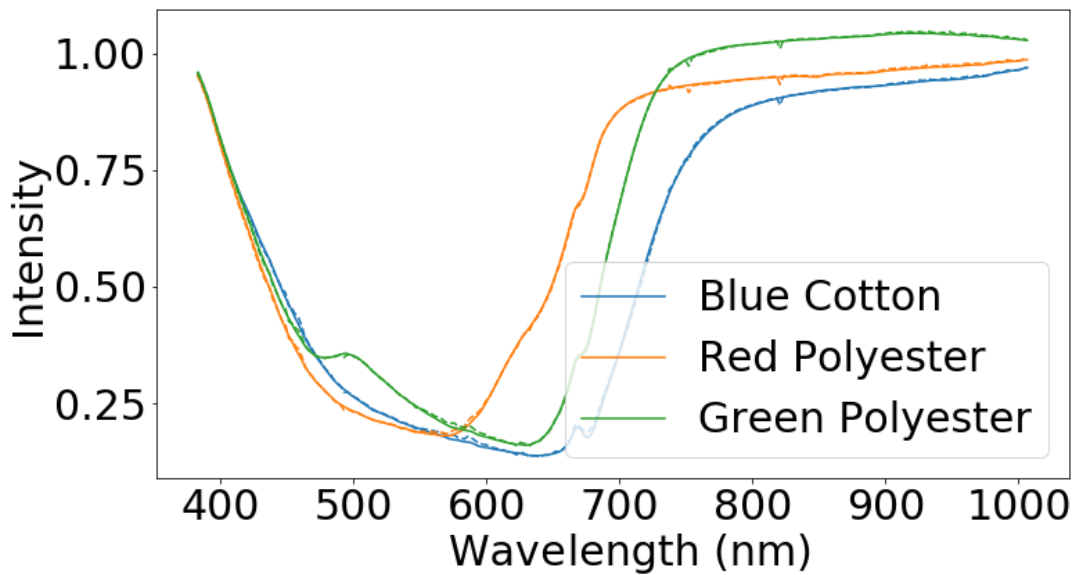
(b)

Figure 5.9: Comparison of selected original spectra and their reconstruction for fabrics. (a) Black fabrics, and (b) coloured fabrics.

SVMs were trained on the original 256 band HSI data and the reconstructed spectra using three, six and eleven channels from the multi-channel system. According to Table 5.2, it can again be seen that the reconstruction using eleven channels generates better
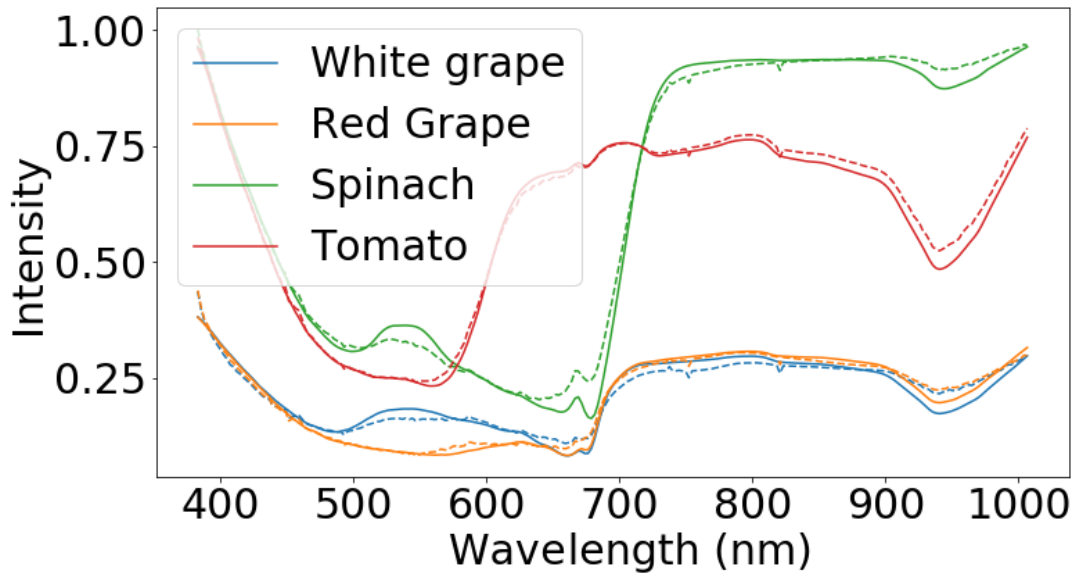
Figure 5.10: Comparison of selected original spectra and their reconstruction for organic
materials.

results than the six-channel one. Specifically, the fabrics show the largest difference,
98.80% with the original data versus 86.84% with three channels, 91.89% with six
channels, and 93.10% with eleven channels. For the organic materials, the classification
accuracy is almost the same for six and eleven channels but almost 7% lower with
three channels. For the Macbeth chart, the classification seems to improve only by a
small margin, even with three channels, a very high accuracy can be achieved. The
decrease in classification accuracy for fabrics is likely attributed to the reduced intensity,
as mentioned above, in the black fabrics. Even though the reconstructions are very
close to the originals, they are almost identical to each other. Consequently, a low
reconstruction error is achieved despite the decreased classification accuracy. When
adding additional channels, subtler variances can be detected and signals with a low
Signal-to-Noise Ratio (SNR) can be better reconstructed. Likewise, for the organic
materials, the surface of the objects is subject to more variation and simple RGB colour
sensing does not seem to account for all the differences in the signals measured with an
HSI system. For the Macbeth chart however, it can be observed that the three RGB
channels produce an almost identical classification after HSI reconstruction, which is
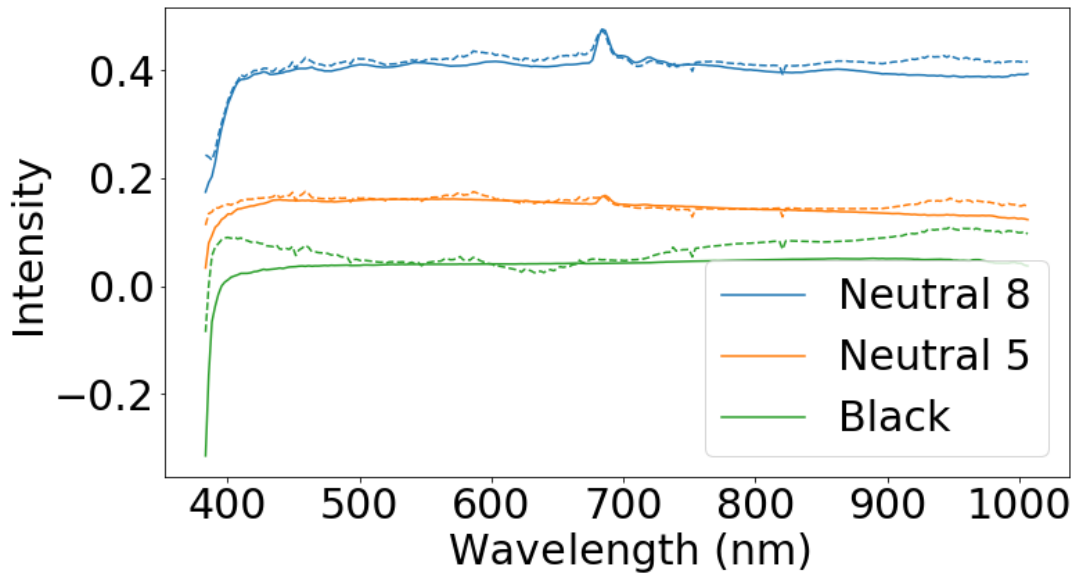
(a)
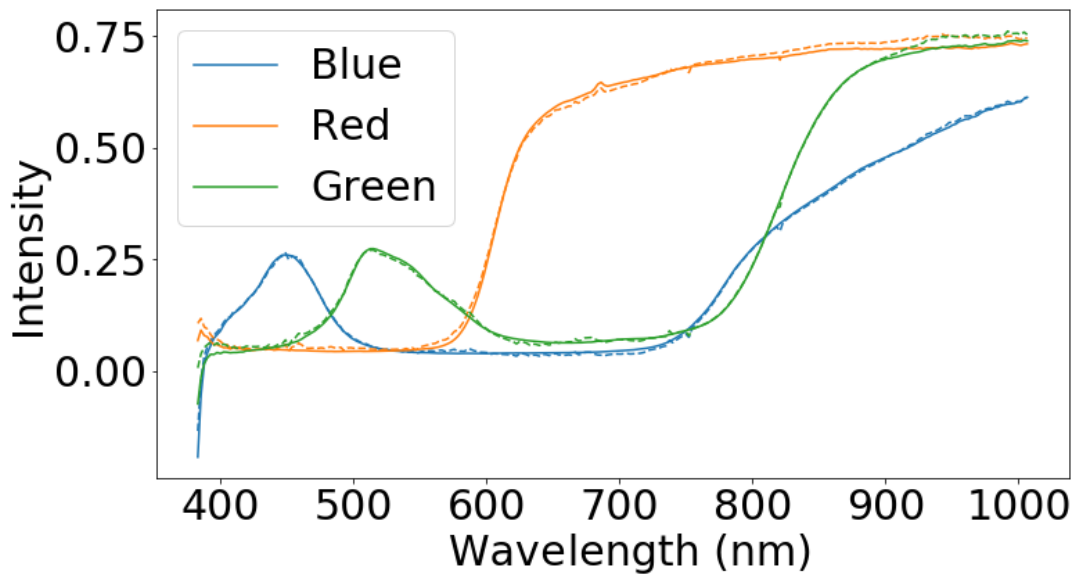
(b)

Figure 5.11: Comparison of selected original spectra and their reconstruction for the Macbeth chart. (a) Coloured patched and (b) black and grey patches.

likely due to the fact that the chart itself is designed to calibrate RGB imaging systems and therefore only displays spectra that are produced by primary colours. Using three RGB channels accounts for almost all variances, where additional channels only help

marginally improving the classification accuracy.

5.4 Summary

In summary, it could be shown that customary and adaptive RGB LED illumination can be integrated with a monochromatic camera to achieve a very cost effective and mobile multi-channel camera system. The training of an MLP that maps multi-channel data to HSI prior is conducted by reconstructing hyperspectral profiles of different objects. Despite RGB LEDs having only three distinct peak wavelengths, mixing them with various intensities allows the generation of sufficient variability to accurately reconstruct the HSI data and therefore adds additional information to the predicative system compared to only three-channel RGB illumination. The physical advantages in using RGB LED technology include its availability and maturity, cost effectiveness, and flexible configuration. Additionally, LED allows ring-shaped (or other arrangement) illumination that enables uniform lighting for each colour channel with no additional measures to eliminate artefacts such as chromatic shadows that are caused by angular non-uniformity. For optimal reconstruction, eleven channels were used with a framerate of 18 images per second and a high spatial pixel resolution. The integration rate can be further increased by sacrificing some of the reconstruction quality. It is also shown that a reconstruction with six channels and a framerate of 33 images per second still generate very good results. The reconstruction is tested on three groups of objects. It shows that a highly accurate reconstruction is possible when the original HSI spectra are sufficiently distinct. Errors occur when the measured signal is low in intensity and with high similarity in the spectral range (400 - 650 nm) covered by the RGB LEDs, regardless of the differences in the spectral region outside. A remedy or improvement is to increase LEDs to cover a broader spectral range. Due to the nature of supervised machine learning, this system is reliant on HSI prior and requires pre-training on spectral signatures of imaged objects to be functional. Given that many HSI applications are very specific and limited in the number of occurring spectra, this poses not a major disadvantage.

Importantly, this prototype system promises a cost effective and mobile hyperspectral imaging setup with a high spatial pixel resolution and framerate. Compared to techniques such as compressive sensing, the proposed system does not require complicated and costly additional optical elements [23] and senses images at a very high spatial resolution. The setup is flexible in configuration and can be trained for various applications. Even though the system is reliant on HSI prior, the HSI data is only needed once to train a system on a specific application, and can then potentially be mass produced for this application. Other systems that rely on spectral prior require additional information on the spectral sensitivity of the sensor [66]. Additionally, the relatively simple design of the MLP allows for an effective reconstruction that can potentially be realised on-board at the system, as opposed to previously proposed systems based on CNNs [67]. The cost-effectiveness and portability are major factors that make this technology interesting for medical and industrial applications. The non-invasive nature of HSI makes it particularly interesting for medical applications and developing specialised imaging devices that are mobile and cost-effective make this a very promising area of future research. Likewise, the food and drink industry increasingly relies on automated quality and safety assurance of their products, requiring technology that is easily deployable and low cost for economically weaker countries. This will in detail be discussed in the following chapter.

Chapter 6

Industrial Case Studies for Food and Drink Quality Monitoring Using Hyperspectral Imaging

6.1 Introduction

The assurance of quality and safety of food and drink products is a global issue of high importance. Food quality monitoring entails many aspects from food safety to quality assurance. In [155], food safety is defined as the probability of not suffering any hazard from consuming the respective product. Food quality in turn defines how well the product is received by the consumer according to several criteria such as appearance and taste. Both are often perceived by consumers to be related, stating that good quality products automatically imply safety [155]. Recent global outbreaks of food related diseases, such as Bovine spongiform encephalopathy (BSE) [156] or salmonella [157], have consistently undermined the trust in the food industry. These diseases can be caused by physical, chemical and biological contaminations, such as the extensive use of artificial chemicals and growth hormones or the outbreak of bacterial infections such as E. coli [158]. An early detection of pathogenic bacteria and other detrimental causes to quality is therefore of high importance for the food industry for

economic, health and humanitarian reasons. HSI can be a particularly valuable tool for the food and drink industry because of its rapid data acquisition and analysis and non-destructive nature. Recent advances in the technology allow for HSI to be used in various spectral regions such as UV [159, 160], NIR [17, 161, 162] and SWIR [163] and in different modalities such as Raman [27] or fluorescence imaging [164]. This versatility has caused a wide range of applications being built in food inspection in the past [165].

Recent applications of HSI in food inspection are vast. In [17], NIR HSI was used to assess the quality of cereal grains in terms of factors such as hardness and chemical composition as well as the safety assurance in terms of detection of fungal and parasitic contamination. Additionally, in [18] it was shown that different types of foodborne bacteria can be distinguished with HSI, highlighting its capability to detect and identify such bacteria. Almond kernels were tested with VIS HSI in [166] for their nitrogen content to determine the protein concentration and therefore ensure quality of the product. The classification of healthy and diseased onions with various techniques was compared in [163] to help sorting in the onion industry. It was established that SWIR HSI performed the best and even outperformed classic NIR spectroscopy. Equally, tea leaves were successfully classified with NIR HSI [161, 162] to help sorting in industry. Another recent focus was put on the quality assessment of meat, such as lamb [167] or beef [168, 169] using VIS and NIR HSI. With regards to physical contamination, the author of this thesis also conducted research to detect and identify contamination on chewing gum [31] in the context of an M.Sc. thesis, where it could be shown that very small foreign objects, such as hair and metal and plastic splinters on the inspected gum could be detected and classified using VIS HSI.

The variety of mentioned application in food inspection covers only a very small percentage of the possibilities enabled by HSI, but it becomes evident that HSI has the potential to simplify industrial processes, increase food safety and deliver better consumer satisfaction. To further stress this point, this chapter will focus on two additional applications. The first study aims at optimising the quality of Scotch Whisky, by determining the expected smokiness of the final product from the barley malt used for distillation. Currently, only distillates can be tested for smokiness and therefore

require blending to achieve a desired level. HSI can potentially purify this process and increase the accuracy. This study was conducted in cooperation with the Scotch Whisky Research Institute (SWRI). A small study is trying to detect bacteria causing listeriosis on salmon fillets. An outbreak of listeriosis, which is a severe type of food poisoning and particularly dangrous for pregnant women [170], in Scotland would be a catastrophe for the Scottish salmon industry and HSI has the potential to help accelerate and improve the current detection methods. The study was conducted in cooperation with Stirling University. Both studies are additional examples of the power of HSI but also highlight limitations in acquiring and analysing the data.

6.2   Estimation of the smokiness of Scotch Whisky

6.2.1   Background

Scotch Whisky is central to the UK economy, accounting for around a quarter of the country's food and drink exports [171]. It is a high-quality spirit drink exclusively produced in Scotland in a manner strictly regulated by law. Maintaining high quality standards during production is therefore of major interest for the industry. Flavour character is central to this, with the flavour compounds present in Scotch coming naturally from the raw materials, and modified or generated through the production process. Certain Scotch Whiskies are characterised by smoky flavours, which are introduced through the exposure of the malted barley to peat smoke during kilning (drying of the malted grain). Volatile phenolic compounds in the peat smoke adhere to the surface of the barley and are carried through the production process giving the smoky character of the final product.

Different distilleries use malted barley that has been peated (smoked) to different degrees, depending on their flavour requirements. Distillers typically specify a set peating level in terms of parts-per-million (ppm) level of phenols, calculated as the total of the major phenolic flavour compounds present on the malt. The current methods to determine the levels of phenols, High Performance Liquid Chromatography (HPLC) or spectrophotometric techniques [172], cannot be carried out on the malt itself. They
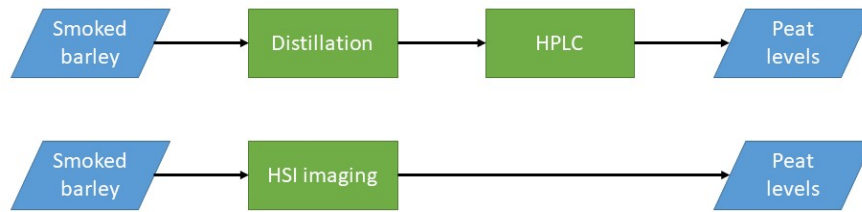
Figure 6.1: Visualisation of the advantages of HSI for the peat level estimation. Traditionally, the barley malt is required to be distilled and HPLC performed to determine the peat levels. These processes require several days and are prone to human error. HSI in comparison can deliver results in real-time from the barley malt and can deliver instant results, prior to any processing of the malt.

rely on a pre-distillation step to extract the phenols, with the analysis performed on the resulting distillate. This is not only time-consuming, but can result in a degree of inaccuracy if the distillation/extraction is not carried out efficiently. For the industry, it would be of significant benefit to be able to determine the phenol levels directly from the malt. Because of the non-destructive nature and the rapid data acquisition, HSI is explored as a potential real-time method. The advantages of HSI to traditional techniques are visualised in Figure 6.1.

In the past, phenols have successfully been detected in seeds, skins and stems of grapes [48, 173] as well as in wood that is used for wine barrels [174] by NIR HSI. Groundwork for the detection of phenols was presented in [175], where NIR HSI covering a spectral range of $950 - 1700$ nm was found to be able to differentiate three different levels of phenol concentrations in peated malt with SVM classification. However, due to the limited number of different concentrations studied, no effective regression model to estimate the exact concentration could be trained. The aim of this study is to extend the work through the evaluation of samples with a wider range of phenol concentrations, scanning them in different spectral regions to further identify the potential of HSI for the estimation and generating more robust systems. To achieve this, ten peated malts with different phenol levels were scanned with a SWIR camera that covers a spectral range of $1000 - 2500$ nm. Additional information is extracted from wavelengths above 1700 nm that have not been explored in the previous study. Phenols also typically show specific absorption characteristics in the UV range with peaks between $260 - 300$ nm [176]. UV

117

hyperspectral imaging is yet quite unexplored with limited applications [159, 160] due to increased difficulties in the imaging process. Not only are UV HSI cameras hardly available and still a subject of research [177,178], broadband light sources in wavelengths under 400 nm with sufficient illuminance are very scarce. In this study, a UV system covering the range of 220 – 400 nm is evaluated for suitability as a novel technology for imaging barley malt and estimating the phenol content. This study focuses only on the use of UV and SWIR HSI, simply because SWIR has proven successful in the past and UV HSI is of interest as it is a yet quite unexplored technology that might show benefits for this particular application. VIS HSI is not considered here as phenolic flavour compounds will unlikely expose any significant absorption bands in this range and it was observed that depending on the type of barley and the dryness, the surface colour varies greatly, which will likely cause significant interference in the VIS range.

6.2.2 Experimental setup

Chemical background

When purchasing malted barley, Scotch Whisky distillers will specify the level of *total phenols* that they require. This is the sum of eight individual phenolic flavour compounds; phenol, guaiacol, p-cresol, m-cresol, o-cresol, 4-methylguaiacol, 4-ethylguaiacol and 4-ethylphenol. Although these may not be the only flavour compounds present in peat smoke, these marker phenols have been used for decades as a measure of smoky character [179, 180]. In today's malt specifications the level of total phenols generally required is in the range of 0 – 50 ppm [172].

Peat can be sourced from various locations across Scotland, including both island and mainland sites. The composition of the peat varies depending on its origin. This influences the relative levels of phenolic compounds on the peated malt, which in turn influences flavour [181]. Each compound imparts a subtly different aroma. Phenol for example tends to be described as medicinal, while guaiacol is experienced as more smoky. The HPLC method, currently used, is both applicable to barley malt steam distillates and new make spirits. The levels are measured in mg/kg or ppm respectively and the total of all compounds combined is used as a marker to the degree of peatiness

of the malted barley. Therefore, in this study, the aim was to estimate the total phenol levels rather than the concentration of the individual compounds.

Data preparation

Some preliminary results of data classification, based on three categories of peating level using NIR HSI were reported in [175]. The objective here was to classify into finer granulated levels with the goal of estimating the actual concentration utilising UV and SWIR HSI technologies. The peated malt samples selected for this study represented the range of total phenol concentrations typically used in Scotch Whisky production. A 124.5 ppm (atypically highly smoked) sample was added to the set to further test the method. The sample set contained in total ten samples with the following total phenol concentrations: 0, 3.8, 8.2, 12.5, 15.5, 20.5, 30, 40, 50 and 124.5 ppm.

During kilning, peat smoke is passed through a bed of malted barley. Due to the nature of this process, each grain has a different amount of phenols adhering to its surface. As a consequence, the concentration cannot be measured from a single spatial point. HPLC can measure the average concentration of a distillate extract, but for image data, this needs to be accounted for. Therefore, the grains were spread out on a flat surface to maximise the surfaces to be imaged using HSI and the mean spectral reflectance of a set of random spatial points was used rather than individual spectra.

Data acquisition

UV imaging was done with the pco. Sensicam UV [182] that covers a spectral range of $220 - 400$ nm with a spectral resolution of 3.8 nm. It requires a specialised illumination with a Hamamatsu L6301-50 Deuterium lamp [183] that has a broadband UV coverage but is very low in light intensity in turn. This results in an increased exposure time of 300 ms. Additionally, 3 x 3 spatial and 4-fold spectral binning were applied. This means that 3 x 3 spatial pixels and 4 spectral bands are added up to not only reduce noise but also to increase the camera's light sensitivity. However, the binning process also degrades the spatial and spectral resolution. This results in an image with 342 pixels per scanned line and 248 active bands. As a result of the increased exposure

time, the UV scanning process requires several minutes, whereas the SWIR imaging
can be done in under a minute. The SWIR imaging was done with the Specim SWIR
system [184] covering a range of $1000 - 2500$ nm with a spectral resolution of 12 nm.
Illumination is done with customary halogen lightbulbs allowing for an exposure time
of only 2.5 ms. Binning was not applied resulting in an image with 384 pixels per line
and 288 active spectral bands.

Figure 6.2 depicts false colour representations of the same barley sample with the
UV camera at 295 nm and Figure 6.3 the data of the SWIR camera at 1483 nm
along with a mean spectrum over the spectral range of each camera of each image
region. Additionally, normalised histograms of both datasets at the same wavelengths
are plotted. One can see that the UV measurements are in general of very low intensity
and expose a much lower SNR. The SWIR range also seems to capture more spectral
features of the barley itself from visual inspection.
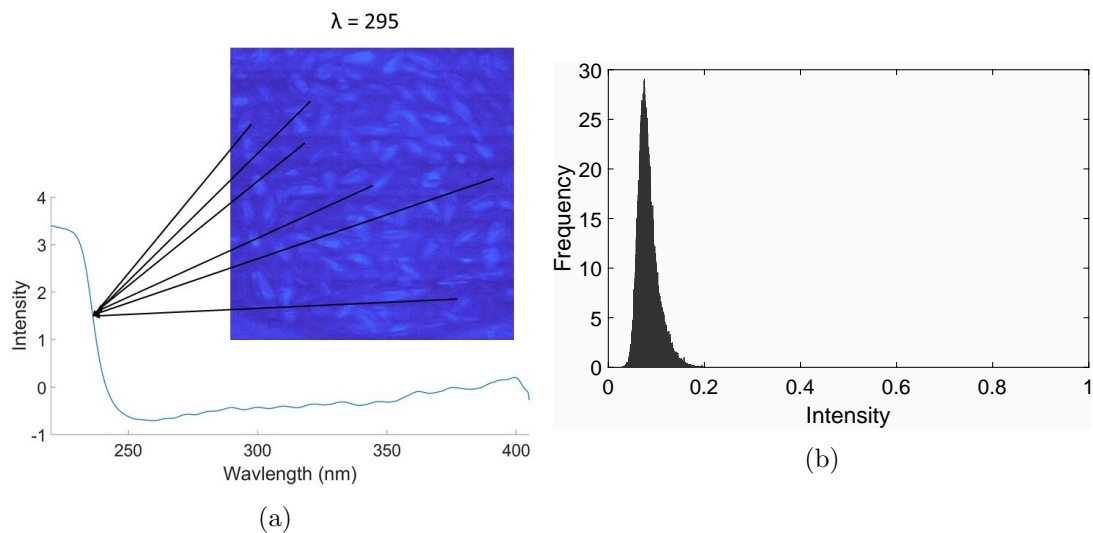
Figure 6.2: Quality of the UV data. (a) Image quality and mean spectral response of
the UV data at the respective wavelength, and (b) according normalised histogram.

Analytical methodology

The spectra used as observations were the pure spectra acquired with both the UV and
SWIR. The UV spectra consist of 248, and the SWIR dataset of 288 spectral bands,
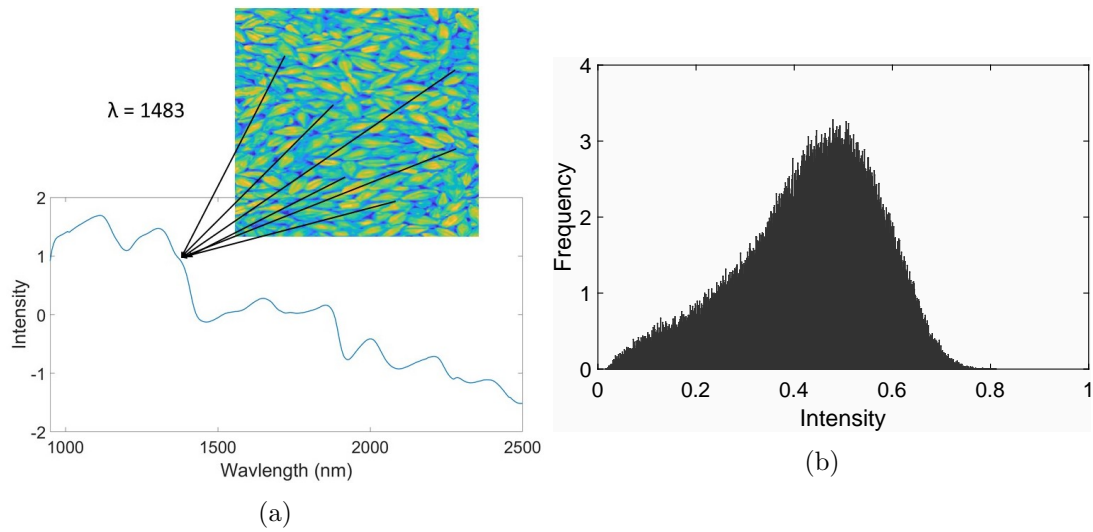
Figure 6.3: Quality of the SWIR data. (a) Image quality and mean spectral response of the SWIR data at the respective wavelength, and (b) according normalised histogram.

which were used in their entirety to maximise the precision of the estimation. To account for shadow and scattering effects due to the very uneven surface of the barley grains, these spectra were converted to SNV.

As the distribution of phenols adhering to the surface of the grains is expected to be very uneven, the pixels will as a consequence have varying spectral responses. In the final application, the mean concentration of an entire batch of barley malt is desired. To achieve this, rather than considering individual pixels as observations, the average over a subset of pixels is taken for analysis instead. To avoid detecting dense regions with locally high concentrations, these subsets are not formed by spatially connected regions but by selecting a subset of pixels randomly distributed over the entire image. Subsets are generated to be disjoint to maximise sample variance.

Two different machine learning approaches were applied for data modelling. As ten different levels of phenol concentration are considered, a classification problem is implied. Classification can help to roughly identify classes of peat concentration which is likely sufficient for the industrial application. For research purposes, it is however of relevance to determine if an exact estimate of the actual concentration can be obtained. Therefore, SVM classification and SVR were employed to predict the exact levels. The purpose of this study is to prove the feasibility of estimating the concentration, which

is why it is only focused on SVM and SVR. Determining the best algorithms for both classification and regression and utilising data reduction techniques should be subject to future investigations.

6.2.3 Experimental results

Classification results under varying settings

To observe the effect of different sizes of pixel subsets, the average over 50, 200 and 500 pixels was taken to generate the observations for classification. Due to different spatial resolutions, this results in a total of 46,702, 11,675 and 4,669 observations for SWIR and 17,174, 4,292 and 1,719 observations for UV in sum for all classes respectively. Out of the observations resulting from averaging 200 pixels in each dataset, 1%, 5% and 10% were then randomly selected for training and the rest for testing the predictive models. Results for both UV and SWIR with varying sizes of subsets and varying ratios of training and testing pixels are shown in Table 6.1. For both datasets, a lower size of pixel subsets also results in a lower classification accuracy. The peat smoking process results in a variable distribution of phenols across the grains. By taking the mean over a certain random subset, it is attempted to introduce a sufficient statistic that can represent the mean concentration of the whole barley batch. The smaller these subsets are, the less likely it is that these subsets represent a sufficient statistic. Larger subsets are therefore more likely to represent a more representative spectrum for the overall concentration and lead to better classification results.

Table 6.1: SVM Classification OA in % for varying number of pixel subset sizes and a training ratio of 5% and varying training ratio with subset size 200

Pixels	UV		SWIR		Ratio	UV		SWIR	
	OA	Kappa	OA	Kappa		OA	Kappa	OA	Kappa
50	65.4	61.6	98.9	98.8	**1%**	76.2	73.6	97.5	97.3
200	90.3	89.2	99.7	99.7	**5%**	90.3	89.3	99.5	99.5
500	96.9	96.7	99.8	99.9	**10%**	93.7	93.0	99.8	99.8

For a subset size of 50 pixels, c.a. 46,000 samples for SWIR and 17,000 for UV were gained. With a higher amount of testing samples, misclassifications are more likely, which also explains a lower OA. This becomes evident particularly for the UV dataset, where a subset size of 50 pixels with 5% training ratio can only achieve a 65.4% OA. Increasing the size can however achieve an OA of 96.9%. The SWIR dataset consistently outperforms the UV dataset with a minimum OA of 98.9% for a subset size of 50 and reaches an OA of 99.8% with increased training ratio or number of pixels. Varying the ratio of training and testing data has a similar effect where only 1% of all samples for the UV dataset can achieve an OA of 76.2% and it reaches 97.5% for the SWIR dataset.

To gain a more detailed insight, confusion matrices for the best and the worst combinations of Table 6.1 are visualised in Figures 6.4 and 6.5 with varying subset sizes and training ratios respectively. In the UV dataset, one can see that with both fixed subset size and training ratio, Classes 1 and 10 seem to be the easiest to discriminate as almost no samples are misclassified. This is likely due to the fact that they represent the extreme cases of 0 and 124.5 ppm phenol concentration. Classes 2 and 5 – 9 seem to have the most misclassified samples as can be observed in Figures 6.4a, 6.4b, 6.5a and 6.5b. For the SWIR dataset, classes 1 – 5 seem to have the most misclassifications, whereas classes 6 and 7 are consistently well classified.

Results from mean spectra of spatially coherent regions

In addition to the classification results obtained under various training ratios and randomly generated subsets, the classification was also tested on mean spectra of spatially coherent regions. For this, classifiers were trained for both datasets using 20% of the samples generated by taking the average over 50 random pixels. 50 pixels were chosen as this generates a larger number of training samples and will likely capture more statistical variations. Validation sets were produced by calculating the average of 50×50 windows for all samples of both datasets. This results in 300 samples for the UV dataset and 840 samples for the SWIR dataset respectively over all classes. Greyscale representations of wavelength 1483 nm and 284 nm for the SWIR and UV datasets respectively are illustrated alongside a ground truth and the classification results in
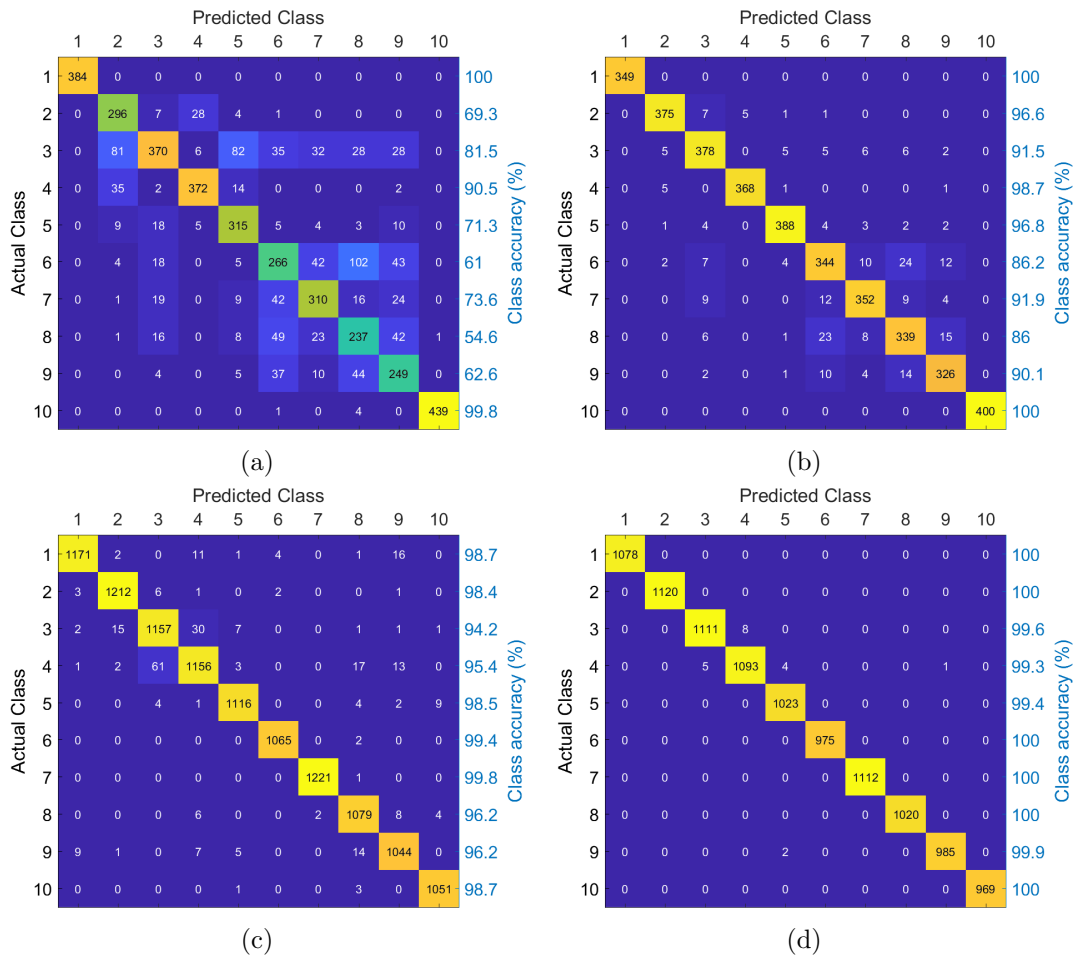
Figure 6.4: Confusion matrices for different subset sizes. (a) UV dataset, 5% training ratio and subset size of 50 pixel, (b) UV dataset, 5% training ratio and subset size of 500 pixel, (c) SWIR dataset, 5% training ratio and subset size of 50 pixel, and (d) SWIR dataset, 5% training ratio and subset size of 500 pixel.

Figures 6.6 and 6.7. For the UV data, classes 1 and 10 are classified best, which is consistent with the results from Figures 6.4 and 6.5. Likewise, classes 6 and 7 are visually classified the best in the SWIR dataset, which again confirms the previous results. The OAs achieved are in both cases much lower than in Table 6.1, particularly for the UV dataset, which is only 47.7%.

As previously stated, the phenol concentrations are very likely to vary drastically for each barley grain. This also results in spatial agglomerations of very high phenol concentration or a total absence of phenols. By taking the mean of spatially coher-
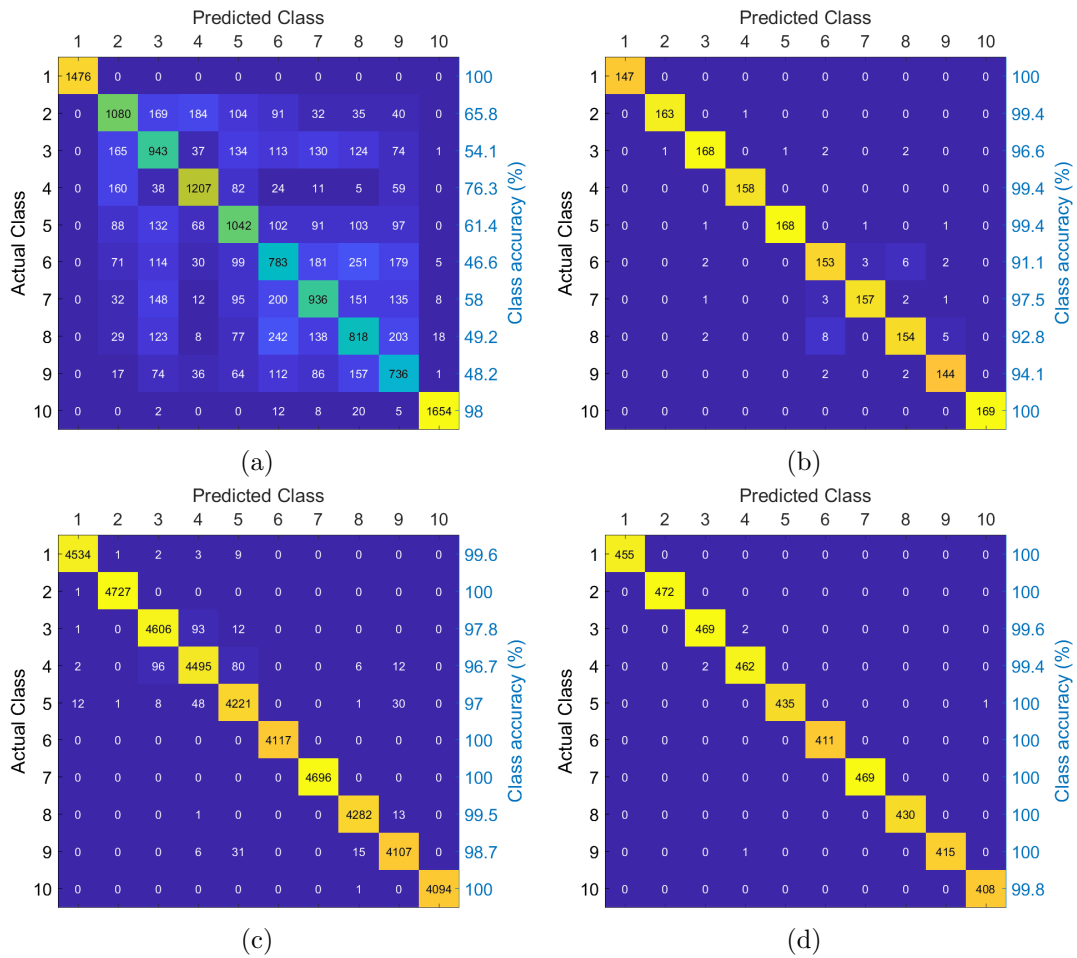
Figure 6.5: Confusion matrices for different training ratios. (a) UV dataset, subset size of 200 pixel and training ratio of 1%, (b) UV dataset, subset size of 200 pixel and training ratio of 10%, (c) SWIR dataset, subset size of 200 pixel and training ratio of 1%, and (d) SWIR dataset, subset size of 200 pixel and training ratio of 10%.

ent regions, samples that represent exactly these variations are likely generated. The ground truth is not necessarily valid anymore as a result. The low classification accuracy is therefore not a reliable statistic. In fact, misclassified pixels can even be interpreted positively as the classifier is able to pick up spatial variations of concentrations. Generating a ground truth to evaluate the correctness is hardly possible which is why different methods of validation are needed. The high accuracies on the randomly generated samples indicate that this classification is quite accurate. With this, it might be able to estimate the spatial distribution of phenol concentration within a barley
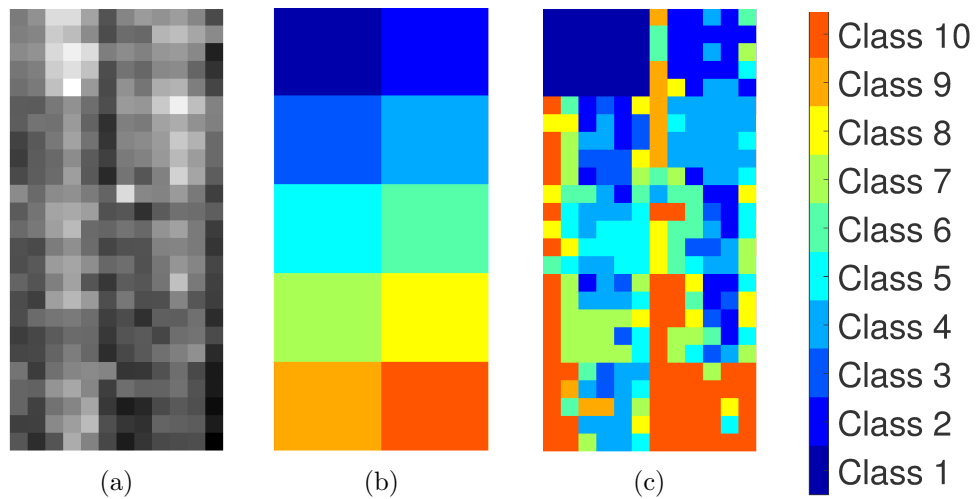
Figure 6.6: Classification of UV. (a) Generated mean pixel at wavelength 295 nm, (b) ground truth, and (c) classification results. OA: 47.7%

batch, providing potentially useful additional information for the industry.

Results of regression

Even though the differentiation between ten concentration levels implies a classification problem, from a scientific point of view, it would also be interesting to estimate the actual concentration level. For this purpose, SVR on the same datasets was applied. Likewise, three regression runs with three selections of random subsets were performed for comparison. Results for varying numbers of pixel subsets and varying ratios of training and testing samples are summarised in Table 6.2. Similar to the classification, it can be seen that the UV data produces significantly worse results especially for small sizes of pixel subsets and resulting higher number of samples and for a low training sample ratio. On the other hand, the SWIR dataset is able to achieve an r^2 value of 0.92 in the worst case, or 0.99 in the best case in comparison to 0.91 from the UV dataset. As mentioned in Section 6.2.2, the working range of phenol concentrations for the whisky production is between 0 - 50 ppm, and the industry is usually interested in only three categories [175]. The achieved RMSE between 0.75 and 0.32 ppm can be interpreted as an average estimation error of 0.75 or 0.32 ppm and therefore is able to estimate the phenol concentration to an acceptable precision. For more accurate levels

(a)   (b)   (c)
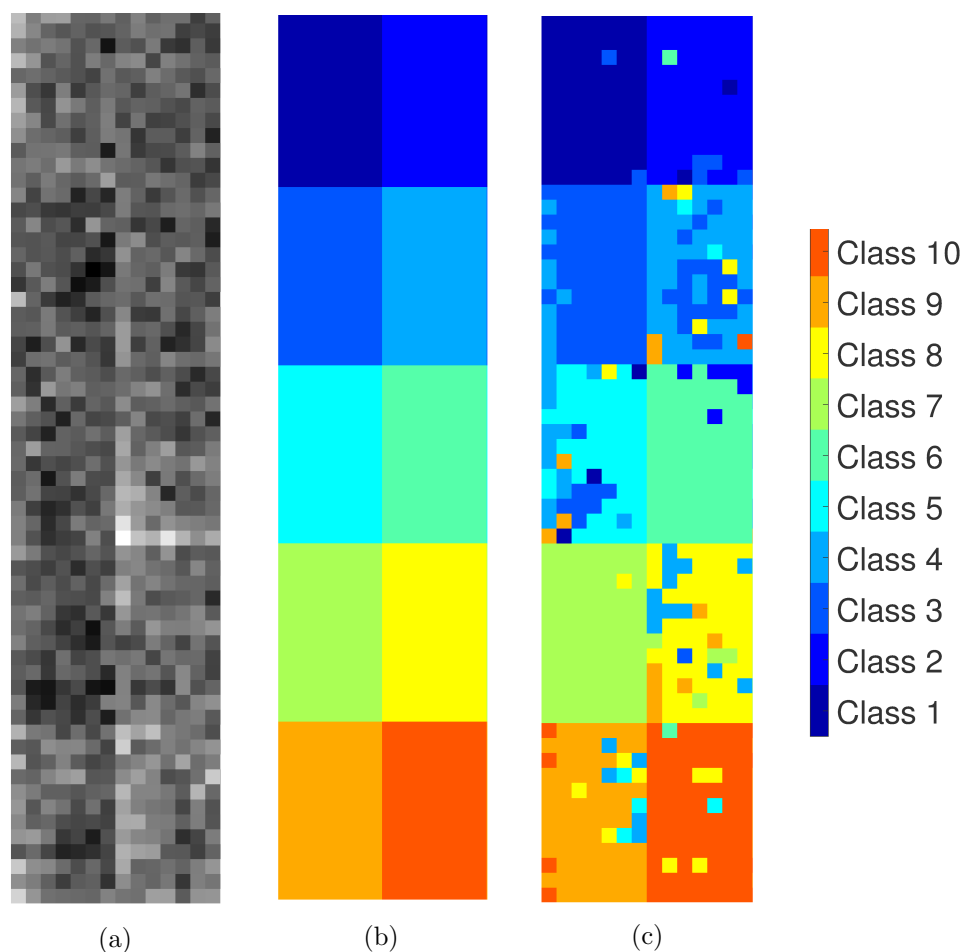
Figure 6.7: Classification of UV data. (a) Generated mean pixel at wavelength 1483 nm, (b) ground truth, and (c) classification results. OA : 84.6%

of the individual compounds, HPLC and other chemometric analysis tools are still a better choice. However, it is important to bear in mind that these compounds are only markers of flavour. As a result, the SWIR HSI may in fact be giving a better measure of all compounds that are of sensory importance.

6.2.4   Summary

In this study, the potential of HSI for the estimation of the concentration of phenolic flavour compounds on malted barley was examined. For this purpose, barley samples

Table 6.2: Correlation coefficient and RMSE of SVR with varying numbers of subset sizes and a training ratio of 5% and varying training ratios with subset size 200

Pixels	UV		SWIR		Ratio	UV		SWIR	
	r^2	RMSE	r^2	RMSE		r^2	RMSE	r^2	RMSE
50	0.56	1.55	0.92	0.75	**1%**	0.74	1.23	0.95	0.63
200	0.82	1.04	0.98	0.42	**5%**	0.83	1.05	0.98	0.43
500	0.91	0.77	0.99	0.32	**10%**	0.85	1.00	0.98	0.36

with ten different concentrations were generated and imaged with a UV HSI system covering a spectral range from 220 – 400 nm and a SWIR system covering a range of 1000 – 2500 nm. The samples were then classified using an SVM with Gaussian RBF kernel. Results show that especially the SWIR dataset is able to discriminate very precisely between ten concentration levels with an OA of up to 99.8%. The UV dataset in comparison performs worse in all cases. Similar results were observed for the SVR, where the UV dataset can only achieve an r^2 value of 0.91 whereas the SWIR dataset can achieve a value up to 0.99 with an RMSE 0.32 ppm. It was also established that UV HSI is a still underdeveloped technology, where the availability of appropriate light sources is scarce which results in aggravated imaging conditions and very low SNR images. In summary, it could be shown that SWIR HSI has in its current stage of development very much the potential to be used in industrial applications to quantify and classify the mean phenol concentration levels of an entire barley batch.

6.3   Detection of Listeria monocytogenes on smoked salmon

6.3.1   Background

Smoked salmon as a ready-to-eat food product is considered a risk product for human listeriosis [185] which is caused by Listeria monocytogenes (L.mono). According to [186], fish products have the highest prevalence of L.mono, especially smoked salmon with a contamination rate of 4.31%. When purchased in vacuum packaging, the fish usually has a shelf-life in refrigerated temperatures of several weeks. As the bacteria

can grow in the absence of oxygen and at temperatures from 1°C to 45°C, they can easily reproduce under normal storage conditions [187]. Even though L.mono has been detected in fresh and seawater, the concentrations are hardly ever high enough to cause concern of potential human infections [186]. The bacterium is however capable of forming very resistant biofilms on working surfaces, which is why the food processing chain is identified as the main source for listeria contamination in hazardous concentrations. A hygienic environment during the processing of the raw and smoked salmon is therefore crucial to avoiding any contamination [185]. The mere detection of the presence is already enough to cause a recall of the product and is therefore to be avoided in any case.

Common ways to detect L.mono involve sampling the food product and enriching growth in the laboratory before enumeration on selective agar [188]. These methods are time consuming and can take up to several days. If listeria contamination can be detected on a batch of salmon fillets, the product may already be in circulation and needs to be recalled. A rapid detection of the pathogen would be of enormous advantage to the industry. Detection of listeria on frankfurters in the short-wave infra-red has been successfully shown in the past [189]. The study presented here tries to identify if the presence of listeria can be detected on various surfaces via HSI.

6.3.2 Experimental setup

The camera used here operates between the NIR and SWIR ranging from 950 - 1700 nm. Three different types of samples were screened. A first attempt was to determine the detectability of the bacteria in a saline solution of sodium chloride at 0.85%. Therefore three different concentrations were produced, 10^8 (high), 10^6 (medium) and 10^4 colony forming units per ml (cfu/ml). A second set of samples screened the viable bacterial grow in colonies on agar. The attempt is to obtain pure spectra from these colonies. A third set of samples attempts to detect the bacteria on salmon fillets. The fillets were inoculated with a high concentration of L.mono and incubated for zero, one and three days at 10°C.

All samples were radiometric calibration and converted to SNV according to Chap-
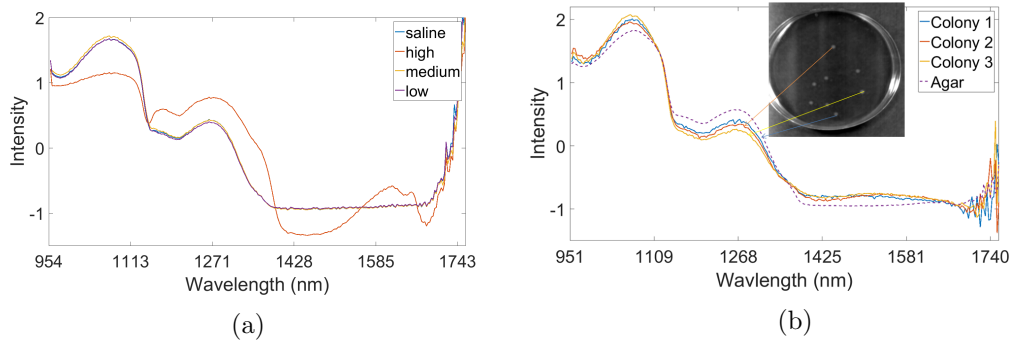
Figure 6.8: Spectra of listeria. (a) Saline and listeria at various concentrations, and (b) bacterial colonies compared to spectrum of agar.

ter 2.2.3 to make the data comparable.

6.3.3 Experimental results

It was at first attempted to determine the detectability of the bacteria in a saline solution. Figure 6.8a shows the measured mean spectra of the solutions with different amounts of L.mono. One can see that the high concentration shows a significantly distinct spectrum especially between 1150 - 1200 nm and 1400 - 1700 nm, but medium and low are almost identical to the pure saline. To mitigate interference of water absorption cause by the saline solution, L.mono colonies on agar were imaged and spectra are compared in Figure 6.8b. In general, a very similar shape of the spectra among the samples can be observed, while the agar seems to expose slight differences in magnitude in certain wavelength areas. The colonies show a slight increase in the region of 1400 - 1700 nm, which would be in accordance with the observations of highly concentrated listeria in saline solution. It is assumed that the colonies are not completely opaque in the NIR and that therefore many mixed spectra of bacteria and agar are measured. The limited spatial resolution of the camera adds to that.

The inoculated salmon fillets show different types of tissue on the measured surface and depending on the individual part of the fillet, also vary in composition of those. As shown in Figure 6.9a, the samples consist of flesh and fat parts to different extents and both supposedly have different spectral signatures. From visual inspection of the spectra in Figure 6.9, it can be seen that there is no consistent profile observable between

130
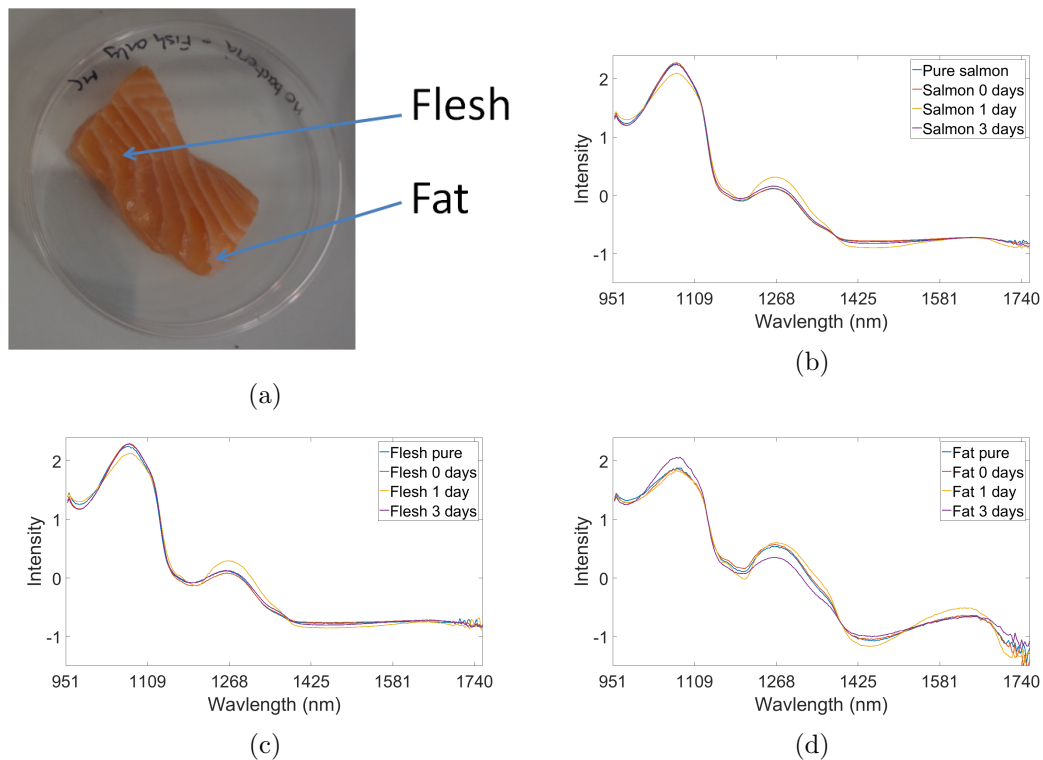
Figure 6.9: Spectra of different parts of the fish. (a) Depiction of fillet with fat and flesh tissue, (b) mean spectra of regions on the whole fillet, (c) spectra of flesh tissue, and (d) spectra of fat tissue.

fillets that were inoculated and the pure salmon fillet. The only difference that seems to prevail is a slight variation of the spectral signature for the one day old fillet in the fleshy part and the whole fillet. Since the three day old fillet does not expose a similar variation, this is likely not to be attributed to listeria contamination. The spectra of the fat parts differ a bit. In wavelengths less than 1400 nm, the three day old fish varies, whereas in wavelengths greater than 1400 nm the one day old differs again. The fact that the samples are from different pieces of the fillet is most likely the explanation for the measured spectral differences. Especially the fat parts seem to vary the most and be the least consistent. From these observations, it is concluded that there are too many unknown factors and too little controlled experimentation to derive any useful information at this stage.

6.3.4   Summary

This study served the purpose of determining whether listeria can be detected by means of NIR HSI. It could be determined that pure listeria colonies show a spectral profile that can well be detected with NIR HSI, especially in the range of 1150 - 1200 nm and 1400 - 1700 nm. The bacterium needs to be present in very high concentrations to be detected, which in real life rarely occurs. As for the salmon fillets, there are many questions yet to be answered. As we are working with an organic organism that comes on its own with a certain bacterial culture on its surface, it is very hard to control the growth of the bacteria. The undefined bacterial composition of the surface and various side effects of the bacterial growth, such as the the production of lipids, can obscure the measured presence of listeria. The value of this study is to lay the cornerstone for future investigations and raise many questions that need to be examined in detail and therefore help future approaches of similar studies.

6.4   Discussion

In this chapter, the applicability of different HSI technologies for food quality monitoring processes was evaluated on two main applications. Based on the specific requirements of the application and the chemical composition of the imaged objects, certain spectral ranges could be determined. Spectroscopic information about chemical or biological substances, such as their absorbance criteria and therefore spectral characteristics, are often available in the context of chemometry. The first study, where the phenol content of barley malt was estimated, showed very accurate classification and regression results. The very specific nature of the available samples, i.e. the fact that phenolic compounds are very unevenly distributed over the samples, required additional preparation of the data to generate observations useful for machine learning. With this pre-processing step, good classification of levels of phenol concentration and a very accurate estimation of the actual levels was achieved. The UV HSI technology utilised here showed potential but also in the current stage of development the limitations of the technology, as sensing and illumination hardware are both open to improvement.

SWIR HSI in contrast proved to be very applicable, underlining its importance for many applications in the field of food quality monitoring. The detection of listeria in the NIR region showed initially promising results, as a quite specific spectral signature could be obtained from a saline solution with a very high concentration of listeria and individual colonies on agar showed similar characteristics. On salmon fillets however, it was quickly established that the rich organic surface of the fish fillets together with a mixture of different tissue material yields very inconsistent data that is hardly usable for machine learning. This highlights a limitation of HSI for food quality monitoring, as experiments need to be performed in highly controllable environments to obtain pure spectra and as a result generate reliably predictive data.

From these two studies, we can clearly derive the difficulties arising, when analysing food products with HSI. Even though HSI has the significant advantage of adding spatial data to spectral data, its quality is often obliterated by inaccurate measurements at pixel level because of effects such as spectral mixing and a low spatial resolution as well as limitations in the spectral capacities of the sensor and dispersive elements. Even after calibration, the measurements obtained often do not always display clearly distinguishable spectral features of the imaged objects but rather a combination of spectra caused by illumination, sensor sensitivity, limited spatial resolution and limitations of the optical components. This causes shortcomings in comparison to very accurate measurements with spectroscopes. Under controlled circumstances, HSI data however works very well in conjunction with appropriate machine learning algorithms. HSI produces a vast number of measurements on pixel level capturing more statistical variability than individual spectroscopic measurements. Machine learning requires input data with sufficient variability to be able to train an informed model. Consequently, not only the spatial mapping of spectra but also the mere amount of measurements highlight the advantages of HSI compared to traditional spectroscopic and chemometric techniques. The rapid data acquisition and analysis capabilities as well as the non-destructive nature of the imaging process are additional factors for an increasing industrial interest. This interest is a motor for the development of low-cost devices, such as the proposed system in Chapter 5, which can be trained and applied in in-

dustrial applications and have the potential to provide precise, portable, and highly time and cost-effective solutions for an industry relying more and more on automated processes.

Chapter 7

Conclusions and Future Work

7.1 Conclusion

The main objective of this thesis was motivated by the rapid development and rise in demand for cost-effective and easy-to-use HSI technology in various commercial and industrial fields, with a particular focus on food quality assessment. This issue was addressed from three different perspectives. At first, HSI data reduction techniques were analysed with a focus on retaining physical attributes of spectral information. Two band selection approaches were developed where the first defines a criterion for the optimal subset that is subsequently optimised, whereas the second approach embeds the identification of the most relevant bands into the learning process of a spectral reconstruction algorithm. Moving on from band selection, it was then focused on effective acquisition of HSI data. For this purpose, a camera system was prototyped that can reconstruct HSI data using deep learning techniques enabling cost-effective, portable and easy data acquisition. Lastly, the capabilities of estimating the presence of concentration of chemical compounds on food products, barley malt in particular, was assessed with respect to optimal data processing and the implementation in a production environment, to provide a context for the developed data reduction and acquisition techniques.

This chapter will draw some conclusions from the presented work and the impact of the developed techniques will be evaluated. Ultimately, ideas for further improvement

and future perspectives will be offered.

1. In Chapter 4, two approaches to hyperspectral band selection were developed. The first algorithm aimed at optimising the MIMR criterion that tries to maximise the Shannon entropy of individual bands of a subset while simultaneously minimising the MI, i.e. the redundant information, between them. It was established that especially for the calculation of the MI, the estimation of the joint PDFs of spectral bands using KDE was necessary to achieve high accuracies. This estimation however comes with an immense computational burden, which is why VarBWFastMI was developed. It significantly reduces the time required to compute the MI by avoiding repeated calculations and also incorporates variable bandwidths for the different spectral bands.

   Based on the improved pre-calculation of the lookup tables to evaluate the MIMR criterion, DGSA was modified for band selection by encoding band subsets as agents in an N-dimensional space. A neighbourhood and movement concept for band subsets based on the Manhattan distance was developed. The algorithm was then benchmarked against different EAs, including the original CSA, PSO, and FA. It was shown that DGSA performs very similar to CSA in terms of classification accuracy, whereas both are outperformed by PSO and FA on specific datasets in terms of speed and accuracy. However, both PSO and FA owe this to their inferior optimisation capabilities of the MIMR algorithm, which results in a much less robust selection quality and exposes shortcomings of the criterion. It could be shown that DGSA performs the most robust in average over different datasets and is easiest to handle because it has only three hyperparameters.

   The second approach focused on the usage of AEs for embedded band selection. The idea to train an AE to reconstruct an input signal and imposing a sparsity constraint on the input weights was extended to HSI. A simple architecture with one hidden layer with equally as many nodes as the input and output was sufficient to achieve accurate reconstruction. Using $L^{2,1}$-regularisation on the input weights during training helps identifying the bands most relevant to the recon-

136

struction. Additionally, the spectral region of the input data was divided into several logical segments and an independent AE was trained on each of these segments. The concept of DD was introduced as a measure of how much information of each segment is required to distribute the number of bands among the segments strategically. This enforces faster convergence of the AEs and a better coverage of the whole spectral range, ultimately leading to competing performance to other state-of-the-art algorithms.

Both band selection algorithms display good techniques to reduce the original data amount to a much lower number of bands (about 50) at the cost of minor decrease in classification accuracy.

2. Chapter 5 introduced a prototype for a cost-effective, portable HSI system. RGB LED enabled time-multiplexed illumination was realised synchronising image integration with the switching between up to eleven different colour settings. A relatively high framerate of the camera allows for video-rate integration of a minimum of 18 up to 200 images per second, depending on the number of channels used. Based on the acquired data, an MLP was designed to reconstruct fully spectrally resolved 256 band data from the multi-channel input given adequate HSI prior. The system was tested on the Macbeth ColourChecker chart to test the reconstruction of spectra of various coloured surfaces and on fabrics of different materials but similar colour shades to prove the capabilities of dealing with subtle colour changes and different materials having different spectral signatures. The reconstruction was also tested on some organic materials, highlighting its capabilities to be able to pick up spectral features such as the chlorophyll red edge. Critically, it was possible to show that adding additional colour channels generated by RGB light sources to the mere RGB image data equally adds value to the reconstruction. It was concluded that given sufficient discriminability between the limited spectra present, the system can be trained for specific applications. Its value derives from the ease of use, dynamic configuration, cost-effectiveness and portability. In the current state, the system is does not have the computa-

tional power to run neural networks. Additional hardware can be added in future versions to the SBC to allow for an on-board reconstruction of the data.

3. Lastly, in Chapter 6, two industrial case studies using various techniques of HSI to determine the concentration of biological and chemical substances were conducted. At first, it was attempted to detect the presence of L.mono on salmon fillets, where it was concluded that while the bacterium can be detected isolated in a very high concentration, the surface of a salmon fillet exposes too many interfering factors, such as proteins, lipids and other bacteria, that the measurements were inconclusive in the conducted study.

   On the other hand, determining the predicted smokiness of Scotch Whisky from barley malt with HSI could be proven to be feasible. Various batches of malt with different concentrations of phenolic flavour compounds that are responsible for the smoky taste in the final product were imaged using UV and SWIR HSI. It was established that the distribution of these phenols is non–uniform across the surface due to the process of smoking, which is inherently hard to control. Generating observations by averaging over random spatial pixel sets in the whole batch helped compensate for these effects. SVM for classification of ten different phenol levels and SVR for the accurate estimation of the actual concentration were both employed successfully. While the results for UV HSI data were only slightly inferior to SWIR, it was shown that UV data acquisition is more complicated due to scarce availability of adequate light sources and the underdeveloped state of the sensing technology. SWIR imaging however is a well established technology posing the potential for this approach to be implemented in a production environment.

From the presented results, it can be concluded that on a larger scale, hyperspectral band selection has proven to be a useful tool in reducing the amount of data. The reduction is however not enough to justify the according construction of multispectral devices, because the high number of bands makes it difficult to produce according systems. Band selection is therefore more efficient as a data reduction on high dimensional input data. As a result, the development of more efficient imaging systems relies on

novel technologies and algorithms that are able to produce similarly informative information to HSI. The capabilities of HSI in the field of food quality monitoring alone, including all its advantages such as data acquisition speed and non-destructive analysis justify the intense research focus in this area.

7.2   Future work

From the presented results and conclusions, some areas were identified that can be expanded in future work. These can be summarised as follows:

1. Based on the results presented from the proposed MIMR-DGSA algorithm, it was clear that it shows some advantages over other, similar heuristics. As it was concluded that the MIMR criterion does not necessarily optimise classification accuracy, a first approach would be to verify this. This could be done by finding alternative datasets with different applications, such as regression, target detection or segmentation. Comparing the optimisation algorithms with respect to their capabilities of each application might shed more light on shortcomings and advantages of the MIMR criterion. Even though DGSA has proven to be more time efficient than CSA, a major bottleneck is still the initialisation step. Improving this step by developing a faster approach than the proposed greedy search could have significant impact on the runtime. Developing a better initialisation strategy also opens possibilities to eliminate a further parameter, the number of candidates n_c, which would ease the use of the algorithm even further. Likewise, the local search performed after each movement operation has potential for improvement, possibly generating better exploration capabilities and ultimately a faster convergence. Another approach for future work could be to combine band selection with feature extraction techniques. In [190], it was shown that prior spatial feature extraction can significantly increase the quality. Techniques such as CNNs have in the past proven very efficient in generating features for image data. This might be utilised here to incorporate spatial information into the spectral features and thereby utilise the HSI data more efficiently. Adjacent pixels

have a high likelihood of showing similar spectral signatures and can therefore help reduce noise.

2. With respect to S-AEBS, it became quite clear that the proposed algorithm has varying effects on different datasets. While AEBS combined with spectral segmentation on the Indian Pines dataset outperformed all other compared algorithms, the selection quality on the Salinas and Pavia University dataset were almost identical to a random selection under the same segmentation. This shows that the definition of AEBS has potential for improvement. One issue is enforcing better sparsity of the input weights. While the additional weight decay regularisation improves the reconstruction results, it should be analysed in detail how much it adversely affects the development of sparsity.

   As pointed out in Chapter 5, the algorithm also lacks the capability of evaluating the correlation between the bands of a selected subset. In the current sate, all incoming neural connection into a node are weighted and added up. Developing an aggregation function that penalises correlation between incoming features might help produce more uncorrelated band subsets with better classification performance.

   A further open question is the spectral segmentation. While the current algorithm performs manual selection of the spectral regions by visually inspecting the correlation matrix, an automated procedure determining the optimal segmentation would be desirable for an easy-to-use algorithm without much prior knowledge about the data constitution.

3. The proposed camera prototype utilises RGB LED illumination in its current configuration. While it was successfully demonstrated that using several colour channels can lead to a very accurate reconstruction, it was also shown that objects with spectral differences outside the range covered by RGB illumination pose limitations for the system. Increasing the spectral range might be of significant benefit here, as LED illumination in the NIR is available.

   Equally, focusing on the development of customised illumination that utilises

LEDs with distinct spectral peaks is very likely going to increase the reconstruction accuracy, because more subtle spectral features present between the red, green and blue channels will be detected more accurately. Determining the number and spectral peaks and their distance will be subject to further investigation.

The proposed reconstruction algorithm also leaves plenty of room for improvement. While the current design aims at architectural simplicity to prove a concept, a more elaborate design incorporating recent advances in deep learning, such as CNNs has huge potential for improvement in this area.

4. As highlighted in this thesis, the L.mono detection on salmon fillets delivered very inconclusive results. To help improve this, experiments need to be performed in a much more controlled environment. This might include washing of the fillets prior to inoculation to mitigate interference of other organisms as well as a careful selection of the pieces of the fillet so they are identical between the samples. Variances in bacterial growth also need to be addressed by increasing the number of samples, and it should be established which concentration of bacteria on the fillet is in reality to be expected on the fish, to determine detectability. It also needs to be further investigated, if flesh and fat sections on the fillet can deliver equally good data or if they require separate consideration. So far, the experiments have only been conducted on fresh salmon fillets. Smoked fillets are produced to increase shelf life among other things. This might lead to data that is more consistent over time and therefore better to detect. Additionally, different levels of maturity of the fish as it is being fished and different types of salmon will increase the variability between the potential samples and require more robust detection systems. These are just a few considerations when approaching further work in this project and highlight the complexities faced when conducting such experiments.

5. The estimation of the smokiness of Scotch Whisky was in the current setup very successful. To move this application into a production environment, more considerations need to be made. First of all, the experiments were all done on barley

malt from one malting company. Different maltsters will have different barley and a carefully trained model on one type of barley will likely not work on a different type. Training a model on a broader sample range is therefore mandatory. Equally, the constitution of the barley in the production chain needs to be considered. Is it going to have the same level of dryness/moisture content and what other adversarial effects will be present that can hamper the imaging process. A stable prediction model is required to deliver reliable results in different environments, with different source material.

References

[1] Jet Propulsion Laboratory, California Institute of Technology. (2019) Airborne Visible/Infrared Imaging Spectrometer. Accessed: 31/01/2019. [Online]. Available: https://aviris.jpl.nasa.gov/index.html

[2] P. A. Mitchell, "Hyperspectral digital imagery collection experiment (HYDICE)," in *Geographic Information Systems, Photogrammetry, and Geological/Geophysical Remote Sensing*, vol. 2587, Paris, France, 1995, pp. 70–95.

[3] S. Holzwarth, A. Müller, M. Habermeyer, R. Richter, A. Hausold, S. Thiemann, and P. Strobl, "HySens - DAIS 7915 / ROSIS Imaging Spectrometers at DLR," in *3rd EARSeL Workshop on Imaging Spectroscopy*, Herrsching, Germany, 2003, p. 12.

[4] B. Koetz, F. Morsdorf, S. van der Linden, T. Curt, and B. Allgöwer, "Multi-source land cover classification for forest fire management based on imaging spectrometry and LiDAR data," *Forest Ecology and Management*, vol. 256, no. 3, pp. 263–271, 2008.

[5] A. Young, S. Marshall, and A. Gray, "Outlier and target detection in aerial hyperspectral imagery: a comparison of traditional and percentage occupancy hit or miss transform techniques," in *Automatic Target Recognition XXVI*, Baltimore, Maryland, United States, 2016, p. 98440S.

[6] R. C. Gonzalez and R. E. Woods, *Digital image processing*. Upper Saddle River, New Jersey: Prentice Hall, 2008.

[7] T. Qiao, J. Ren, M. Sun, J. Zheng, and S. Marshall, "Effective compression of hyperspectral imagery using an improved 3D DCT approach for land-cover analysis in remote-sensing applications," *International Journal of Remote Sensing*, vol. 35, no. 20, pp. 7316–7337, 2014.

[8] J. Zabalza, J. Ren, J. Zheng, H. Zhao, C. Qing, Z. Yang, P. Du, and S. Marshall, "Novel segmented stacked autoencoder for effective dimensionality reduction and feature extraction in hyperspectral imaging," *Neurocomputing*, vol. 185, pp. 1–10, 2016.

[9] J. Zabalza, J. Ren, M. Yang, Y. Zhang, J. Wang, S. Marshall, and J. Han, "Novel Folded-PCA for improved feature extraction and data reduction with hyperspectral imaging and SAR in remote sensing," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 93, pp. 112–122, 2014.

[10] T. Qiao, Z. Yang, J. Ren, P. Yuen, H. Zhao, G. Sun, S. Marshall, and J. A. Benediktsson, "Joint bilateral filtering and spectral similarity-based sparse representation: A generic framework for effective feature extraction and data classification in hyperspectral imaging," *Pattern Recognition*, vol. 77, pp. 316–328, 2018.

[11] J. Zabalza, J. Ren, J. Zheng, J. Han, H. Zhao, S. Li, and S. Marshall, "Novel two-dimensional singular spectrum analysis for effective feature extraction and data classification in hyperspectral imaging," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 53, no. 8, pp. 4418–4433, 2015.

[12] G. Taskin, H. Kaya, and L. Bruzzone, "Feature selection based on high dimensional model representation for hyperspectral images," *IEEE Transactions on Image Processing*, vol. 26, no. 6, pp. 2918–2928, 2017.

[13] J. Feng, L. Jiao, F. Liu, T. Sun, and X. Zhang, "Unsupervised feature selection based on maximum information and minimum redundancy for hyperspectral images," *Pattern Recognition*, vol. 51, pp. 295–309, 2016.

[14] M. E. Pawlowski, J. G. Dwight, T.-U. Nguyen, and T. S. Tkaczyk, "High performance image mapping spectrometer (IMS) for snapshot hyperspectral imaging applications," *Optics Express*, vol. 27, no. 2, p. 1597, 2019.

[15] C. Deng, X. Hu, J. Suo, Y. Zhang, Z. Zhang, and Q. Dai, "Snapshot hyperspectral imaging via spectral basis multiplexing in Fourier domain," *Optics Express*, vol. 26, no. 25, p. 32509, 2018.

[16] N. R. Gomer, C. W. Gardner, and M. P. Nelson, "Handheld and mobile hyperspectral imaging sensors for wide-area standoff detection of explosives and chemical warfare agents," in *Next-Generation Spectroscopic Technologies IX*, Baltimore, Maryland, United States, 2016, p. 98550M.

[17] K. Sendin, P. J. Williams, and M. Manley, "Near infrared hyperspectral imaging in quality and safety evaluation of cereals," *Critical Reviews in Food Science and Nutrition*, vol. 58, no. 4, pp. 575–590, 2018.

[18] T.-L. Kammies, M. Manley, P. A. Gouws, and P. J. Williams, "Differentiation of foodborne bacteria using NIR hyperspectral imaging and multivariate data analysis," *Applied Microbiology and Biotechnology*, vol. 100, no. 21, pp. 9305–9320, 2016.

[19] P. Qiu, A. J. Gentles, and S. K. Plevritis, "Fast calculation of pairwise mutual information for gene regulatory network reconstruction," *Computer Methods and Programs in Biomedicine*, vol. 94, no. 2, pp. 177–180, 2009.

[20] M. B. Dowlatshahi, H. Nezamabadi-Pour, and M. Mashinchi, "A discrete gravitational search algorithm for solving combinatorial optimization problems," *Information Sciences*, vol. 258, pp. 94–107, 2014.

[21] H. Aasen, E. Honkavaara, A. Lucieer, and P. J. Zarco-Tejada, "Quantitative remote sensing at ultra-high resolution with UAV spectroscopy: A review of sensor technology, measurement procedures, and data correctionworkflows," *Remote Sensing*, vol. 10, no. 7, p. 1091, 2018.

[22] D.-W. Sun, *Hyperspectral imaging for food quality analysis and control.* London: Academic Press, 2010.

[23] G. R. Arce, D. J. Brady, L. Carin, H. Arguello, and D. S. Kittle, "Compressive Coded Aperture Spectral Imaging: An Introduction," *IEEE Signal Processing Magazine*, vol. 31, no. 1, pp. 105–115, 2014.

[24] Y. August, C. Vachman, Y. Rivenson, and A. Stern, "Compressive hyperspectral imaging by random separable projections in both the spatial and the spectral domains," *Applied Optics*, vol. 52, no. 10, p. D46, 2013.

[25] C. Li, T. Sun, K. F. Kelly, and Y. Zhang, "A compressive sensing and unmixing scheme for hyperspectral data processing," *IEEE Transactions on Image Processing*, vol. 21, no. 3, pp. 1200–1210, 2012.

[26] X. Lin, G. Wetzstein, Y. Liu, and Q. Dai, "Dual-coded compressive hyperspectral imaging," *Optics Letters*, vol. 39, no. 7, p. 2044, 2014.

[27] J. V. Thompson, J. N. Bixler, B. H. Hokr, G. D. Noojin, M. O. Scully, and V. V. Yakovlev, "Single-shot chemical detection and identification with compressed hyperspectral Raman imaging," *Optics Letters*, vol. 42, no. 11, p. 2169, 2017.

[28] N. Hagen and M. W. Kudenov, "Review of snapshot spectral imaging technologies," *Optical Engineering*, vol. 52, no. 9, p. 090901, 2013.

[29] A. Young, "Atmospheric Correction and Anomaly Detection of Remotely Sensed Hyperspectral Imagery," Ph.D. dissertation, Department of Electronic and Electrical Engineering, University of Strathclyde, Glasgow, UK, 2017.

[30] H. Yao and D. Lewis, "Spectral Preprocessing and Calibration Techniques," in *Hyperspectral Imaging for Food Quality Analysis and Control*, D.-W. Sun, Ed. London: Academic Press, 2010, pp. 45–78.

[31] J. Tschannerl, "Effective Acquisition, Transmission, and Analysis of Hyperspectral Imagery Using Mobile Enabled Embedded Platforms," Master's thesis, De-

partment of Computer Science, University of Applied Sciences, Munich, Germany, 2016.

[32] Å. Rinnan, F. van den Berg, and S. B. Engelsen, "Review of the most common pre-processing techniques for near-infrared spectra," *TrAC - Trends in Analytical Chemistry*, vol. 28, no. 10, pp. 1201–1222, 2009.

[33] M. L. Amodio, I. Capotorto, M. M. A. Chaudhry, and G. Colelli, "The use of hyperspectral imaging to predict the distribution of internal constituents and to classify edible fennel heads based on the harvest time," *Computers and Electronics in Agriculture*, vol. 134, pp. 1–10, 2017.

[34] H. Yu, H. Liu, N. Wang, Y. Yang, A. Shi, L. Liu, H. Hu, R. Mzimbiri, and Q. Wang, "Rapid and visual measurement of fat content in peanuts by using the hyperspectral imaging technique with chemometrics," *Analytical Methods*, vol. 8, no. 41, pp. 7482–7492, 2016.

[35] G. F. Hughes, "On the Mean Accuracy of Statistical Pattern Recognizers," *IEEE Transactions on Information Theory*, vol. 14, no. 1, pp. 55–63, 1968.

[36] J. Zabalza, J. Ren, J. Ren, Z. Liu, and S. Marshall, "Structured covariance principal component analysis for real-time onsite feature extraction and dimensionality reduction in hyperspectral imaging," *Applied Optics*, vol. 53, no. 20, p. 4440, 2014.

[37] P. Bajorski, "Application of nonnegative principal component analysis in hyperspectral imaging," in *Imaging Spectrometry XI*, San Diego, California, United States, 2006, p. 63020G.

[38] X. Jia and J. A. Richards, "Segmented principal components transformation for efficient hyperspectral remote-sensing image display and classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 37, no. 1 II, pp. 538–542, 1999.

[39] P. Mishra, C. B. Cordella, D. N. Rutledge, P. Barreiro, J. M. Roger, and B. Diezma, "Application of independent components analysis with the JADE algorithm and NIR hyperspectral imaging for revealing food adulteration," *Journal of Food Engineering*, vol. 168, pp. 7–15, 2016.

[40] A. Hyvärinen and E. Oja, "Independent component analysis: Algorithms and applications," *Neural Networks*, vol. 13, no. 4-5, pp. 411–430, 2000.

[41] C.-I. Chang, *Hyperspectral Remote Sensing: Techniques for Spectral Detection and Classification.* New York: Kluwer Academic/Plenum Publishers, 2003.

[42] T. Qiao, J. Ren, Z. Wang, J. Zabalza, M. Sun, H. Zhao, S. Li, J. A. Benediktsson, Q. Dai, and S. Marshall, "Effective Denoising and Classification of Hyperspectral Images Using Curvelet Transform and Singular Spectrum Analysis," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 55, no. 1, pp. 119–133, 2017.

[43] Y. Saeys, I. Inza, and P. Larrañaga, "A review of feature selection techniques in bioinformatics," *Bioinformatics*, vol. 23, no. 19, pp. 2507–2517, 2007.

[44] H. G. Vijouyeh and G. Taskin, "A comprehensive evaluation of feature selection algorithms in hyperspectral image classification," in *2016 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, Beijing, China, 2016, pp. 489–492.

[45] Y. Liu, Y. Chen, K. Tan, H. Xie, L. Wang, X. Yan, W. Xie, and Z. Xu, "Maximum relevance, minimum redundancy band selection based on neighborhood rough set for hyperspectral data classification," *Measurement Science and Technology*, vol. 27, no. 12, p. 125501, 2016.

[46] P. Bajcsy and P. Groves, "Methodology for Hyperspectral Band Selection," *Photogrammetric Engineering & Remote Sensing*, vol. 70, no. 7, pp. 793–802, 2004.

[47] D. Yang and W. Bao, "Group Lasso-Based Band Selection for Hyperspectral Image Classification," *IEEE Geoscience and Remote Sensing Letters*, vol. 14, no. 12, pp. 2438–2442, 2017.

[48] N. Zhang, X. Liu, X. Jin, C. Li, X. Wu, S. Yang, J. Ning, and P. Yanne, "Determination of total iron-reactive phenolics, anthocyanins and tannins in wine grapes of skins and seeds based on near-infrared hyperspectral imaging," *Food Chemistry*, vol. 237, pp. 811–817, 2017.

[49] X. Luo, R. Xue, and J. Yin, "Information-Assisted Density Peak Index for Hyperspectral Band Selection," *IEEE Geoscience and Remote Sensing Letters*, vol. 14, no. 10, pp. 1870–1874, 2017.

[50] Q. Du and H. Yang, "Similarity-Based Unsupervised Band Selection for Hyperspectral Image Analysis," *IEEE Geoscience and Remote Sensing Letters*, vol. 5, no. 4, pp. 564–568, Oct 2008.

[51] E. Amaldi and V. Kann, "On the approximability of minimizing nonzero variables or unsatisfied relations in linear systems," *Theoretical Computer Science*, vol. 209, no. 1, pp. 237–260, 1998.

[52] Y. Liu, G. Wang, H. Chen, H. Dong, X. Zhu, and S. Wang, "An improved particle swarm optimization for feature selection," *Journal of Bionic Engineering*, vol. 8, no. 2, pp. 191–200, 2011.

[53] B. Xue, M. Zhang, and W. N. Browne, "Particle swarm optimization for feature selection in classification: A multi-objective approach," *IEEE Transactions on Cybernetics*, vol. 43, no. 6, pp. 1656–1671, 2013.

[54] E. Emary, H. M. Zawbaa, K. K. A. Ghany, A. E. Hassanien, and B. Parv, "Firefly Optimization Algorithm for Feature Selection," in *Proceedings of the 7th Balkan Conference on Informatics Conference - BCI '15*. New York, New York, USA: ACM Press, 2015, pp. 1–7.

[55] H. Su, Y. Cai, and Q. Du, "Firefly-Algorithm-Inspired Framework with Band Selection and Extreme Learning Machine for Hyperspectral Image Classification," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 10, no. 1, pp. 309–320, 2017.

[56] H. Su, Q. Du, G. Chen, and P. Du, "Optimized hyperspectral band selection using particle swarm optimization," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 7, no. 6, pp. 2659–2670, 2014.

[57] H. Su, B. Yong, and Q. Du, "Hyperspectral band selection using improved firefly algorithm," *IEEE Geoscience and Remote Sensing Letters*, vol. 13, no. 1, pp. 68–72, 2016.

[58] A. E. Eiben and S. K. Smit, "Evolutionary Algorithm Parameters and Methods to Tune Them," in *Autonomous Search*.   Berlin, Heidelberg: Springer Berlin Heidelberg, 2011, pp. 15–36.

[59] Y. Zhan, D. Hu, H. Xing, and X. Yu, "Hyperspectral Band Selection Based on Deep Convolutional Neural Network and Distance Density," *IEEE Geoscience and Remote Sensing Letters*, vol. 14, no. 12, pp. 2365–2369, 2017.

[60] X. Cao, C. Wei, J. Han, and L. Jiao, "Hyperspectral Band Selection Using Improved Classification Map," *IEEE Geoscience and Remote Sensing Letters*, vol. 14, no. 11, pp. 2147–2151, 2017.

[61] B. Chandra and R. K. Sharma, "Exploring autoencoders for unsupervised feature selection," in *2015 International Joint Conference on Neural Networks (IJCNN)*, Killarney, Ireland, 2015, pp. 1–6.

[62] K. Han, C. Li, and X. Shi, "Autoencoder inspired unsupervised feature selection," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Calgary, Alberta, Canada, 2018.

[63] Y. Bian, Q. Liu, Z. Zhang, D. Liu, A. Hussian, C. Kuang, H. Li, and X. Liu, "Portable multi-spectral lens-less microscope with wavelength-self-calibrating imaging sensor," *Optics and Lasers in Engineering*, vol. 111, pp. 25–33, 2018.

[64] M. Goel, S. N. Patel, E. Whitmire, A. Mariakakis, T. S. Saponas, N. Joshi, D. Morris, B. Guenter, M. Gavriliu, and G. Borriello, "HyperCam: Hyperspectral Imaging for Ubiquitous Computing Applicaitons," in *2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing UbiComp '15*. Osaka, Japan: ACM Press, 2015, pp. 145–156.

[65] J. Herrera-Ramírez, M. Vilaseca, and J. Pujol, "Portable multispectral imaging system based on light-emitting diodes for spectral recovery from 370 to 1630 nm," *Applied Optics*, vol. 53, no. 14, p. 3131, 2014.

[66] B. Arad and O. Ben-Shahar, "Sparse recovery of hyperspectral signal from natural RGB images," in *The 14th European Conference on Computer Vision*, 2016, pp. 19–34.

[67] S. Galliani, C. Lanaras, D. Marmanis, E. Baltsavias, and K. Schindler, "Learned spectral super-resolution," *CoRR*, vol. abs/1703.09470, 2017. [Online]. Available: http://arxiv.org/abs/1703.09470

[68] Y. B. Can and R. Timofte, "An efficient CNN for spectral reconstruction from RGB images," *CoRR*, vol. abs/1804.04647, 2018. [Online]. Available: http://arxiv.org/abs/1804.04647

[69] G. Ding, Y. Guo, K. Chen, C. Chu, J. Han, and Q. Dai, "DECODE: Deep Confidence Network for Robust Image Classification," *IEEE Transactions on Image Processing*, pp. 1–1, 2019.

[70] C. Chen, J. Han, J. Liu, B. Zhang, and S. Luan, "Gabor Convolutional Networks," *IEEE Transactions on Image Processing*, vol. 27, no. 9, pp. 4357–4366, 2018.

[71] C.-I. Chang, *Hyperspectral Data Processing: Algorithm Design and Analysis*. Hoboken, NJ, USA: John Wiley & Sons, Inc., 2013.

[72] A. F. H. Goetz, "Three decades of hyperspectral remote sensing of the Earth: A personal view," *Remote Sensing of Environment*, vol. 113, no. SUPPL. 1, pp. S5–S16, 2009.

[73] A. F. Goetz, G. Vane, J. E. Solomon, and B. N. Rock, "Imaging spectrometry for earth remote sensing," *Science*, vol. 228, no. 4704, pp. 1147–1153, 1985.

[74] Z. Lee, K. L. Carder, S. K. Hawes, R. G. Steward, T. G. Peacock, and C. O. Davis, "Model for the interpretation of hyperspectral remote-sensing reflectance," *Applied Optics*, vol. 33, no. 24, p. 5721, 1994.

[75] K. L. Carder, P. Reinersman, R. F. Chen, F. Muller-Karger, C. O. Davis, and M. Hamilton, "AVIRIS calibration and application in coastal oceanic environments," *Remote Sensing of Environment*, vol. 44, no. 2-3, pp. 205–216, 1993.

[76] S. L. Ustin, D. A. Roberts, J. A. Gamon, G. P. Asner, and R. O. Green, "Using Imaging Spectroscopy to Study Ecosystem Processes and Properties," *BioScience*, vol. 54, no. 6, p. 523, 2004.

[77] W. S. Lee, V. Alchanatis, C. Yang, M. Hirafuji, D. Moshou, and C. Li, "Sensing technologies for precision specialty crop production," *Computers and Electronics in Agriculture*, vol. 74, no. 1, pp. 2–33, 2010.

[78] E. Honkavaara, H. Saari, J. Kaivosoja, I. Pölönen, T. Hakala, P. Litkey, J. Mäkynen, and L. Pesonen, "Processing and assessment of spectrometric, stereoscopic imagery collected using a lightweight UAV spectral camera for precision agriculture," *Remote Sensing*, vol. 5, no. 10, pp. 5006–5039, 2013.

[79] B. Vajna, A. Farkas, H. Pataki, Z. Zsigmond, T. Igricz, and G. Marosi, "Testing the performance of pure spectrum resolution from Raman hyperspectral images of differently manufactured pharmaceutical tablets," *Analytica Chimica Acta*, vol. 712, pp. 45–55, 2012.

[80] S. Wilczyński, R. Koprowski, M. Marmion, P. Duda, and B. Błońska-Fajfrowska, "The use of hyperspectral imaging in the VNIR (400–1000 nm) and SWIR range

(1000–2500 nm) for detecting counterfeit drugs with identical API composition," *Talanta*, vol. 160, pp. 1–8, 2016.

[81] M. Sun, D. Zhang, Z. Wang, J. Ren, B. Chai, and J. Sun, "What's Wrong with the Murals at the Mogao Grottoes: A Near-Infrared Hyperspectral Imaging Method." *Scientific reports*, vol. 5, p. 14371, 2015.

[82] A. Polak, T. Kelman, P. Murray, S. Marshall, D. J. Stothard, N. Eastaugh, and F. Eastaugh, "Use of infrared hyperspectral imaging as an aid for paint identification," *Journal of Spectral Imaging*, vol. 5, no. 1, pp. 1–10, 2016.

[83] L. M. Kandpal, J. Tewari, N. Gopinathan, P. Boulas, and B. K. Cho, "In-process control assay of pharmaceutical microtablets using hyperspectral imaging coupled with multivariate analysis," *Analytical Chemistry*, vol. 88, no. 22, pp. 11 055–11 061, 2016.

[84] P. Y. Sacré, P. Lebrun, P. F. Chavez, C. D. Bleye, L. Netchacovitch, E. Rozet, R. Klinkenberg, B. Streel, P. Hubert, and E. Ziemons, "A new criterion to assess distributional homogeneity in hyperspectral images of solid pharmaceutical dosage forms," *Analytica Chimica Acta*, vol. 818, pp. 7–14, 2014.

[85] M. Nathan, A. S. Kabatznik, and A. Mahmood, "Hyperspectral imaging for cancer detection and classification," in *3rd Biennial South African Biomedical Engineering Conference (SAIBMEC 2018)*, Stellenbosch, Sourh Africa, 2018, pp. 1–4.

[86] G. Lu, L. Halig, D. Wang, Z. G. Chen, and B. Fei, "Hyperspectral imaging for cancer surgical margin delineation: registration of hyperspectral and histological images," in *3rd Biennial South African Biomedical Engineering Conference (SAIBMEC 2018)*, Stellenbosch, Sourh Africa, 2014, p. 90360S.

[87] G. Lu and B. Fei, "Medical hyperspectral imaging: a review," *Journal of Biomedical Optics*, vol. 19, no. 1, p. 010901, 2014.

[88] K. Gill, J. Ren, S. Marshall, S. Karthick, and J. Gilchrist, "Quality-assured fingerprint image enhancement and extraction using hyperspectral imaging," in *4th International Conference on Imaging for Crime Detection and Prevention 2011 (ICDP 2011)*, London, UK, 2011, pp. P23–P23.

[89] L. N. Brewer, J. A. Ohlhausen, P. G. Kotula, and J. R. Michael, "Forensic analysis of bioagents by X-ray and TOF-SIMS hyperspectral imaging," *Forensic Science International*, vol. 179, no. 2-3, pp. 98–106, 2008.

[90] S. Marshall, T. Kelman, T. Qiao, P. Murray, and J. Zabalza, "Hyperspectral imaging for food applications," in *23rd European Signal Processing Conference (EUSIPCO 2015)*, Nice, France, 2015, pp. 2854–2858.

[91] T. Qiao, J. Ren, C. Craigie, J. Zabalza, C. Maltin, and S. Marshall, "Singular spectrum analysis for improving hyperspectral imaging based beef eating quality evaluation," *Computers and Electronics in Agriculture*, vol. 115, pp. 21–25, 2015.

[92] C. Cortes and V. Vapnik, "Support-Vector Networks," *Machine Learning*, vol. 20, no. 3, pp. 273–297, 1995.

[93] C. J. C. Burges, "Geometry and Invariance in Kernel Based Methods," in *Advances in Kernel Methods—Support Vector Learning*, B. Scholkopf, Ed. MIT Press, 1998, pp. 89–116.

[94] C.-C. Chang, C.-J. Lin, and T. Tieleman, "LIBSVM : A Library for Support Vector Machines," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 307, pp. 1–39, 2008.

[95] S. O. Crichton, S. M. Kirchner, V. Porley, S. Retz, G. von Gersdorff, O. Hensel, and B. Sturm, "High pH thresholding of beef with VNIR hyperspectral imaging," *Meat Science*, vol. 134, pp. 14–17, 2017.

[96] X. Kang, X. Xiang, S. Li, and J. A. Benediktsson, "PCA-Based Edge-Preserving Features for Hyperspectral Image Classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 55, no. 12, pp. 1–12, 2017.

[97] C.-I. Chang, "Spectral information divergence for hyperspectral image analysis," in *International Geoscience and Remote Sensing Symposium. (IGARSS' 99)*, vol. 1, Hamburg, Germany, 1999, pp. 509–511.

[98] C. E. Shannon, "A Mathematical Theory of Communication," *Bell System Technical Journal*, vol. 27, no. 4, pp. 623–656, 1948.

[99] H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information: Criteria of Max-Dependency, Max-Relevance, and Min-Redundancy," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 8, pp. 1226–1238, 2005.

[100] P. A. Estévez, M. Tesmer, C. A. Perez, and J. M. Zurada, "Normalized mutual information feature selection," *IEEE Transactions on Neural Networks*, vol. 20, no. 2, pp. 189–201, 2009.

[101] N. Kwak and C. H. Choi, "Input feature selection by mutual information based on Parzen window," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 12, pp. 1667–1671, 2002.

[102] M. Bennasar, Y. Hicks, and R. Setchi, "Feature selection using Joint Mutual Information Maximisation," *Expert Systems with Applications*, vol. 42, no. 22, pp. 8520–8532, 2015.

[103] A. Hacine-Gharbi, P. Ravier, R. Harba, and T. Mohamadi, "Low bias histogram-based estimation of mutual information for feature selection," *Pattern Recognition Letters*, vol. 33, no. 10, pp. 1302–1308, 2012.

[104] A. Collignon, F. Maes, D. Delaere, D. Vandermeulen, P. Suetens, and G. Marchal, "Automated multi-modality image registration based on information theory," in *14th International Conference on Information Processing in Medical Imaging*, Ile de Berder, France, 1995, pp. 263–274.

[105] F. Maes, A. Collignon, D. Vandermeulen, G. Marchal, and P. Suetens, "Multimodality image registration by maximization of mutual information." *IEEE Transactions on Medical Imaging*, vol. 16, no. 2, pp. 187–98, 1997.

[106] E. Parzen, "On estimation of a probability density function and mode," *The Annals of Mathematical Statistics*, vol. 33, no. 3, pp. 1065–1076, 1962.

[107] S. T. Chiu, "Bandwidth selection for kernel density estimation," *The Annals of Statistics*, vol. 19, no. 4, pp. 1883 – 1905, 1991.

[108] N. B. Heidenreich, A. Schindler, and S. Sperlich, "Bandwidth selection for kernel density estimation: A review of fully automatic selectors," *AStA Advances in Statistical Analysis*, vol. 97, no. 4, pp. 403–433, 2013.

[109] H. Shimazaki and S. Shinomoto, "Kernel bandwidth optimization in spike rate estimation," *Journal of Computational Neuroscience*, vol. 29, no. 1-2, pp. 171–182, 2010.

[110] P. Ramachandran and T. J. Perkins, "Adaptive bandwidth kernel density estimation for next-generation sequencing data." *BMC proceedings*, vol. 7, no. Suppl 7, p. S7, 2013.

[111] V. Kumar and S. Minz, "Feature Selection: A literature Review," *Smart Computing Review*, vol. 4, no. 3, 2014.

[112] A. E. Eiben and C. A. Schippers, "On evolutionary exploration and exploitation," *Fundamenta Informaticae*, vol. 35, no. 1-4, pp. 35–50, 1998.

[113] A. E. Eiben and J. E. Smith, *Introduction to Evolutionary Computing*, ser. Natural Computing Series. Berlin, Heidelberg: Springer Berlin Heidelberg, 2003.

[114] R. O. Duda, P. E. P. E. Hart, and D. G. Stork, *Pattern classification*, 2nd ed. New York, New York: Wiley, 2001.

[115] B. Karlik, "Performance analysis of various activation functions in generalized MLP architectures of neural networks," *International Journal of Artificial Intelligence and Expert Systems*, vol. 1, no. 4, pp. 111–122, 2015.

[116] B. Ding, H. Qian, and J. Zhou, "Activation functions and their characteristics in deep neural networks," in *Proceedings of the 30th Chinese Control and Decision Conference, CCDC 2018*, Shenyang, China, 2018, pp. 1836–1841.

[117] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," *Nature*, vol. 323, no. 6088, pp. 533–536, 1986.

[118] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. Cambridge, Massachussets: MIT Press, 2016.

[119] R. Tibshiranit, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 58, no. 1, pp. 267–288, 1996.

[120] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," in *Advances in Neural Information Processing Systems (NIPS 2012)*, Lake Tahoe, Nevada, 2012, pp. 1097–1105.

[121] C. C. Tan and C. Eswaran, "Using autoencoders for mammogram compression," *Journal of Medical Systems*, vol. 35, no. 1, pp. 49–58, 2011.

[122] L. Meng, S. Ding, and Y. Xue, "Research on denoising sparse autoencoder," *International Journal of Machine Learning and Cybernetics*, vol. 8, no. 5, pp. 1719–1729, 2017.

[123] J. Su, S. Wu, B. Zhang, C. Wu, Y. Qin, and D. Xiong, "A neural generative autoencoder for bilingual word embeddings," *Information Sciences*, vol. 424, pp. 287–300, 2018.

[124] A. Janecek, W. Gansterer, M. Demel, and G. Ecker, "On the Relationship Between Feature Selection and Classification Accuracy." *Journal of Machine Learning Research*, vol. 4, pp. 90–105, 2008.

[125] A. Cheriyadat and L. M. Bruce, "Why principal component analysis is not an appropriate feature extraction method for hyperspectral data," in *International*

Geoscience and Remote Sensing Symposium (IGARSS 2003), vol. 6, Toulouse, France, 2003, pp. 3420–3422.

[126] J. Miao and L. Niu, "A Survey on Feature Selection," *Procedia Computer Science*, vol. 91, pp. 919–926, 2016.

[127] H. Mühlenbein, M. Gorges-Schleuter, and O. Krämer, "Evolution algorithms in combinatorial optimization," *Parallel Computing*, vol. 7, no. 1, pp. 65–85, 1988.

[128] J. Liu, W. Zhong, and L. Jiao, "A multiagent evolutionary algorithm for combinatorial optimization problems," *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, vol. 40, no. 1, pp. 229–240, 2010.

[129] M. E. Aydin and T. C. Fogarty, "A distributed evolutionary simulated annealing algorithm for combinatorial optimisation problems," *Journal of Heuristics*, vol. 10, no. 3, pp. 269–292, 2004.

[130] L. N. De Castro and F. J. Von Zuben, "Learning and optimization using the clonal selection principle," *IEEE Transactions on Evolutionary Computation*, vol. 6, no. 3, pp. 239–251, jun 2002.

[131] P. Qiu. Fast calculation of pairwise mutual information based on kernel estimation. Accessed: 11/02/2019. [Online]. Available: http://pengqiu.gatech.edu/software/FastPairMI/index.htm

[132] E. Rashedi, H. Nezamabadi-pour, and S. Saryazdi, "GSA: A Gravitational Search Algorithm," *Information Sciences*, vol. 179, no. 13, pp. 2232–2248, 2009.

[133] A. Zhang, G. Sun, J. Ren, X. Li, Z. Wang, and X. Jia, "A Dynamic Neighborhood Learning-Based Gravitational Search Algorithm," *IEEE Transactions on Cybernetics*, vol. 48, no. 1, pp. 436–447, 2018.

[134] G. Sun, P. Ma, J. Ren, A. Zhang, and X. Jia, "A stability constrained adaptive alpha for gravitational search algorithm," *Knowledge-Based Systems*, vol. 139, pp. 200–213, 2018.

[135] S. H. Jacobsen, K. A. Sullivan, and A. W. Johnson, "Discrete Manufacturing Process Design Optimization Using Generalized Hill Climbing Algorithms," *Engineering Optimization*, vol. 31, no. 2, pp. 247–260, 2007.

[136] A. Martínez-Usó, F. Pla, J. M. Sotoca, and P. García-Sevilla, "Clustering-based hyperspectral band selection using information measures," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 45, no. 12, pp. 4158–4171, 2007.

[137] J. C. Dunn, "A Fuzzy Relative of the ISODATA Process and Its Use in Detecting Compact Well-Separated Clusters," *Journal of Cybernetics*, vol. 3, no. 3, pp. 32–57, 1973.

[138] A. Ihler. (2003) Kernel density estimation toolbox for matlab (r13). Accessed: 14/09/2017. [Online]. Available: https://www.ics.uci.edu/~ihler/code/kde.html

[139] B. M. and S. F., "On the complexity of neural network classifiers: A comparison between shallow and deep architectures," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 25, no. 8, pp. 1553–1565, 2014.

[140] A. K. Mahlein, U. Steiner, C. Hillnhütter, H. W. Dehne, and E. C. Oerke, "Hyperspectral imaging for small-scale analysis of symptoms caused by different sugar beet diseases," *Plant Methods*, vol. 8, no. 1, p. 3, 2012.

[141] J. Qin and R. Lu, "Measurement of the absorption and scattering properties of turbid liquid foods using hyperspectral imaging," *Applied Spectroscopy*, vol. 61, no. 4, pp. 388–396, 2007.

[142] C. Yang, J. H. Everitt, M. R. Davis, and C. Mao, "A CCD camera-based hyperspectral imaging system for stationary and airborne applications," *Geocarto International*, vol. 18, no. 2, pp. 71–80, 2003.

[143] T. Wu, L. Zhang, B. Peng, H. Zhang, Z. Chen, and M. Gao, "Real-time progressive hyperspectral remote sensing detection methods for crop pest and diseases," in *Remotely Sensed Data Compression, Communications, and Processing XII*, Baltimore, Maryland, 2016, p. 987410.

159

References

[144] P. Gonzalez, J. Pichette, B. Vereecke, B. Masschelein, A. Lambrechts, L. Kraso-vitski, and L. Bikov, "An extremely compact and high-speed line-scan hyperspec-tral imager covering the SWIR range," in *Image Sensing Technologies: Materials, Devices, Systems, and Applications V*, Orlando, Florida, 2018, p. 19.

[145] S. Janchaysang, S. Sumriddetchkajorn, and P. Buranasiri, "Tunable filter-based multispectral imaging for detection of blood stains on construction material sub-strates Part 1 Developing blood stain discrimination criteria," *Applied Optics*, vol. 51, no. 29, p. 6984, 2012.

[146] B. Tan, N. Liao, L. Tian, J. Wang, and Y. Lianry, "High dynamic range mul-tispectral imaging using liquid crystal tunable filter," in *Optoelectronic Imaging and Multimedia Technology*, Beijing, China, 2010.

[147] L. Pan, H.-C. Li, Y.-J. Sun, and Q. Du, "Hyperspectral Image Reconstruction by Latent Low-Rank Representation for Classification," *IEEE Geoscience and Remote Sensing Letters*, vol. 15, no. 9, pp. 1422–1426, 2018.

[148] Y. Jia, J. He, and Z. Luo, "Compressive Hyperspectral Imaging Reconstruction by Spatial and Spectral Joint Prior," in *Proceedings of the 3rd International Con-ference on Intelligent Information Processing (ICIIP '18)*, Guilin, China, 2018, pp. 135–140.

[149] J. Sevilla, G. Martín, and J. M. P. Nascimento, "Parallel hyperspectral image re-construction using random projections," in *High-Performance Computing in Geo-science and Remote Sensing VI*, Edinburgh, United Kingdom, 2016, p. 1000707.

[150] Basler AG. (2019) acA1300-200um - Basler ace. Accessed: 19/03/2019. [Online]. Available: https://www.baslerweb.com/en/products/cameras/area-scan-cameras/ace/aca1300-200um/

[151] Arduino. (2019) Arduino Uno Rev3. Accessed: 19/03/2019. [Online]. Available: https://store.arduino.cc/arduino-uno-rev3

References

[152] Adafruit Industries, LLC. (2019) NeoPixel Ring - 24 x 5050 RGB LED with Integrated Drivers. Accessed: 19/03/2019. [Online]. Available: https://www.adafruit.com/product/1586

[153] ODroid UK. (2019) Odroid XU4Q including PSU. Accessed: 19/03/2019. [Online]. Available: https://www.odroid.co.uk/hardkernel-odroid-xu4?product_id=813

[154] A. Petrulis, P. Vitta, J. Aglinskaite, R. Vaicekauskas, and A. Åukauskas, "Metameric Light Sources: A Recent Paradigm for Functional Lighting," *Proceedings of the Latvian Academy of Sciences, Section B: Natural, Exact, and Applied Sciences*, vol. 71, no. 5, pp. 366–371, 2017.

[155] W. Van Rijswijk and L. J. Frewer, "Consumer perceptions of food quality and safety and their relation to traceability," *British Food Journal*, vol. 110, no. 10, pp. 1034–1046, 2008.

[156] N. Nathanson, J. Wilesmith, and C. Griot, "Bovine spongiform encephalopathy (BSE): Causes and consequences of a common source epidemic," *American Journal of Epidemiology*, no. 11, pp. 959–969, 1997.

[157] A. Hardy, "Salmonella: A continuing problem," *Postgraduate Medical Journal*, vol. 80, no. 947, pp. 541–545, 2004.

[158] A. Wilcock, M. Pun, J. Khanona, and M. Aung, "Consumer attitudes, knowledge and behaviour: A review of food safety issues," *Trends in Food Science & Technology*, vol. 15, no. 2, pp. 56–66, 2004.

[159] H. Li, H. Lyu, N. Liao, and W. Wu, "Measuring spatially varying, multispectral, ultraviolet bidirectional reflectance distribution function with an imaging spectrometer," *Optical Engineering*, vol. 55, no. 12, p. 124106, 2016.

[160] R. Merkel, "Latent Fingerprint Aging from a Hyperspectral Perspective: First Qualitative Degradation Studies Using UV/VIS Spectroscopy," in *9th Interna-*

tional Conference on IT Security Incident Management and IT Forensics (IMF 2015), Magdeburg, Germany, 2015, pp. 121–135.

[161] T. Kelman, J. Ren, and S. Marshall, "Effective classification of Chinese tea samples in hyperspectral imaging," *Artificial Intelligence Research*, vol. 2, no. 4, 2013.

[162] P. Mishra, A. Nordon, J. Tschannerl, G. Lian, S. Redfern, and S. Marshall, "Near-infrared hyperspectral imaging for non-destructive classification of commercial tea products," *Journal of Food Engineering*, vol. 238, pp. 70–77, 2018.

[163] M. N. Islam, G. Nielsen, S. Stærke, A. Kjær, B. Jørgensen, and M. Edelenbos, "Novel non-destructive quality assessment techniques of onion bulbs: a comparative study," *Journal of Food Science and Technology*, vol. 55, no. 8, pp. 3314–3324, 2018.

[164] M. S. Kim, Y. R. Chen, and P. M. Mehl, "Hyperspectral reflectance and fluorescence imaging system for food quality and safety," *Transactions of the ASAE*, vol. 44, no. 3, pp. 721–729, 2001.

[165] Y. Z. Feng and D. W. Sun, "Application of Hyperspectral Imaging in Food Safety Inspection and Control: A Review," *Critical Reviews in Food Science and Nutrition*, vol. 52, no. 11, pp. 1039–1058, 2012.

[166] T. Gama, H. Wallace, S. Trueman, I. Tahmasbian, and S. Bai, "Hyperspectral imaging for non-destructive prediction of total nitrogen concentration in almond kernels," *Acta Horticulturae*, no. 1219, pp. 259–264, 2018.

[167] T. Qiao, J. Ren, J. Zabalza, and S. Marshall, "Prediction of lamb eating quality using hyperspectral imaging," in *Optical Characterization of Materials (OCM 2015)*, Karlsruhe, Germany, 2015, pp. 15–25.

[168] T. Qiao, J. Ren, C. Craigie, J. Zabalza, C. Maltin, and S. Marshall, "Quantitative Prediction of Beef Quality Using Visible and NIR Spectroscopy with Large Data Samples Under Industry Conditions," *Journal of Applied Spectroscopy*, vol. 82, no. 1, pp. 137–144, 2015.

[169] Y. Liu, D. W. Sun, J. H. Cheng, and Z. Han, "Hyperspectral Imaging Sensing of Changes in Moisture Content and Color of Beef During Microwave Heating Process," *Food Analytical Methods*, vol. 11, no. 9, pp. 2472–2484, 2018.

[170] T. Mateus, J. Silva, R. L. Maia, and P. Teixeira, "Listeriosis during Pregnancy: A Public Health Concern," *ISRN Obstetrics and Gynecology*, vol. 2013, pp. 1–6, 2013.

[171] Scotch Whisky Association. (2015) Facts & Figures about Scotch Whisky. Accessed: 14/09/2017. [Online]. Available: http://www.scotch-whisky.org.uk/what-we-do/facts-figures/

[172] T. Bringhurst and J. Brosnan, "Scotch whisky: raw material selection and processing," in *Whisky*, I. Russell and G. Stewart, Eds. London, United Kingdom: Elsevier, 2014, pp. 49–122.

[173] M. J. Jara-Palacios, F. J. Rodríguez-Pulido, D. Hernanz, M. L. Escudero-Gilete, and F. J. Heredia, "Determination of phenolic substances of seeds, skins and stems from white grape marc by near-infrared hyperspectral imaging," *Australian Journal of Grape and Wine Research*, vol. 22, no. 1, pp. 11–15, 2016.

[174] B. Baca-Bocanegra, J. Nogales-Bueno, J. M. Hernández-Hierro, and F. J. Heredia, "Evaluation of extractable polyphenols released to wine from cooperage byproduct by near infrared hyperspectral imaging," *Food Chemistry*, vol. 244, pp. 206–212, 2018.

[175] J. Tschannerl, J. Ren, F. Jack, S. Marshall, H. Zhao, J. Tschannerl, J. Ren, F. Jack, S. Marshall, and H. Zhao, "Employing NIR-SWIR hyperspectral imaging to predict the smokiness of scotch whisky," in *Optical Characterization of Materials (OCM 2017)*, Karlsruhe, Germany, 2017, pp. 1–12.

[176] J. C. Dearden and W. F. Forbes, "Light Absorption Studies: Part XIV. The Ultraviolet Absorption Spectra of Phenols," *Canadian Journal of Chemistry*, vol. 37, no. 8, pp. 1294–1304, 1959.

[177] P. S. Hsu, D. Lauriola, N. Jiang, J. D. Miller, J. R. Gord, and S. Roy, "Fiber-coupled, UV–SWIR hyperspectral imaging sensor for combustion diagnostics," *Applied Optics*, vol. 56, no. 21, p. 6029, 2017.

[178] M. Zucco, V. Caricato, A. Egidi, and M. Pisani, "A Hyperspectral Camera in the UVA Band," *IEEE Transactions on Instrumentation and Measurement*, vol. 64, no. 6, pp. 1425–1430, 2015.

[179] J. S. Swan and D. Howie, "Sensory and analytical studies on the regional composition of Scotch malt whiskies," in *Current Developments in Mating, Brewing and Distilling*, F. G. Priest and I. Campbell, Eds. London, United Kingdom: The Institute of Brewing, 1982, pp. 129–142.

[180] F. J. Thomson, "The estimation of total phenol on malt as a guide to degree of peating of peated distilling malts," in *Current Developments in Mating, Brewing and Distilling*, F. G. Priest and I. Campbell, Eds. London, United Kingdom: The Institute of Brewing, 1982, pp. 273–278.

[181] B. M. Harrison and F. G. Priest, "Composition of peats used in the preparation of malt for Scotch whisky production - influence of geographical source and extraction depth," *Journal of Agricultural and Food Chemistry*, vol. 57, no. 6, pp. 2385–2391, 2009.

[182] PCO AG. (2008) Sensicam UV. Accessed: 19/03/2019. [Online]. Available: https://www.pco.de/fileadmin/user_upload/db/products/datasheet/BR_SCUVE_20080428.pdf

[183] Hamamatsu Photonics K.K. (2019) Deuterium lamp (L2D2 lamp) L6301-50. Accessed: 19/03/2019. [Online]. Available: https://www.hamamatsu.com/eu/en/product/type/L6301-50/index.html

[184] Specim, Spectral Imaging Ltd. (2019) Spectral Camera SWIR. Accessed: 19/03/2019. [Online]. Available: http://www.specim.fi/downloads/SWIR_SpeCam_ver4-15.pdf

[185] L. M. Rørvik, "Listeria monocytogenes in the smoked salmon industry," *International Journal of Food Microbiology*, vol. 62, no. 3, pp. 183–190, 2000.

[186] D. E. Gombas, Y. Chen, R. S. Clavero, and V. N. Scott, "Survey of Listeria monocytogenes in ready-to-eat foods," *Journal of Food Protection*, vol. 66, no. 4, pp. 559–569, 2003.

[187] M. Favretti, A. Pezzuto, and G. Arcangeli, "Listeria monocytogenes: A Dangerous and Insidious Pathogen in Seafood," in *Microbes in Food and Health*, N. Garg, S. M. Abdel-Aziz, and A. Aeron, Eds. Springer International Publishing, 2016, pp. 333–348.

[188] U. Gasanov, D. Hughes, and P. M. Hansbro, "Methods for the isolation and identification of Listeria spp. and Listeria monocytogenes: A review," *FEMS Microbiology Reviews*, vol. 29, no. 5, pp. 851–875, 2005.

[189] S. Atanassova, H. Daskalov, and T. Stoyanchev, "Application of near-infrared spectroscopy for detection of Listeria monocytogenes in frankfurters." *Scientific Works of the University of Food Technologies - Plovdiv*, vol. 56, no. 1, pp. 561–566, 2009.

[190] L. Jiao, J. Han, X. Cao, B. Ji, L. Wang, and Y. Ji, "Fast hyperspectral band selection based on spatial feature extraction," *Journal of Real-Time Image Processing*, vol. 15, no. 3, pp. 555–564, 2018.

[191] M. F. Baumgardner, L. L. Biehl, and D. A. Landgrebe. (2015) 220 Band AVIRIS Hyperspectral Image Data Set: June 12, 1992 Indian Pine Test Site 3. Accessed: 11/02/2019. [Online]. Available: https://purr.purdue.edu/publications/1947/1

Appendix A

Datasets

A.1 Indian Pines dataset

This dataset was collected by the AVIRIS sensor in 1992 and is a subregion of an image covering the Indian Pines test site in North-western Indiana. It consists of 145×145 pixels and 224 spectral reflectance bands that cover a range from 400 - 2500 nm. It contains two thirds agriculture and one third forest and other vegetation. Some of the agriculture is in early growth phases and therefore only covers a very small percentage. The ground truth is divided in 16 classes with varying amount of samples, as shown in Figure A.1. To reduce noise effects in the data, the water absorption band regions are removed, i.e. bands [104 - 108], [150 - 163] and 220. The data is available at [191].

A.2 Salinas dataset

The Salinas dataset was also collected by the AVIRIS sensor over the Salinas Valley, California. The dataset comprises 512×217 pixels and again 224 bands and has therefore a significantly higher data amount than the Indian Pines scene. The scene depicts vegetables, bare soils, and vineyard fields. The ground truth also contains 16 classes, depicted in Figure A.2, containing mainly agricultural areas. To reduce noise effects in the data, the water absorption band regions were removed, i.e. bands [104 - 108], [150 - 163] and 220.

Stone-Steel-Towers (93)
Buildings-Grass-Trees-Drives (386)
Woods (1265)
Wheat (205)
Soybean-clean (593)
Soybean-mintill (2455)
Soybean-notill (972)
Oats (20)
Hay-windrowed (478)
Grass-pasture-mowed (28)
Grass-trees (730)
Grass-pasture (483)
Corn (237)
Corn-mintill (830)
Corn-notill (1428)
Alfalfa (46)

(a)   (b)

Figure A.1: Indian Pines dataset. (a) Pseudo RGB representation and (b) ground truth. The number of samples per class are indicated in brackets behind the name.

A.3   Pavia University dataset

The Pavia University dataset was acquired by the ROSIS over Pavia, Northern Italy. It consists of 610×610 samples with a spatial resolution of 1.3 m per pixel. Of the original 115 spectral bands covering a range within 430 nm - 860 nm, 12 noisy bands were removed, leaving 103 bands. As shown in Figure A.3, the ground truth contains 9 classes including urban, vegetation and soil.
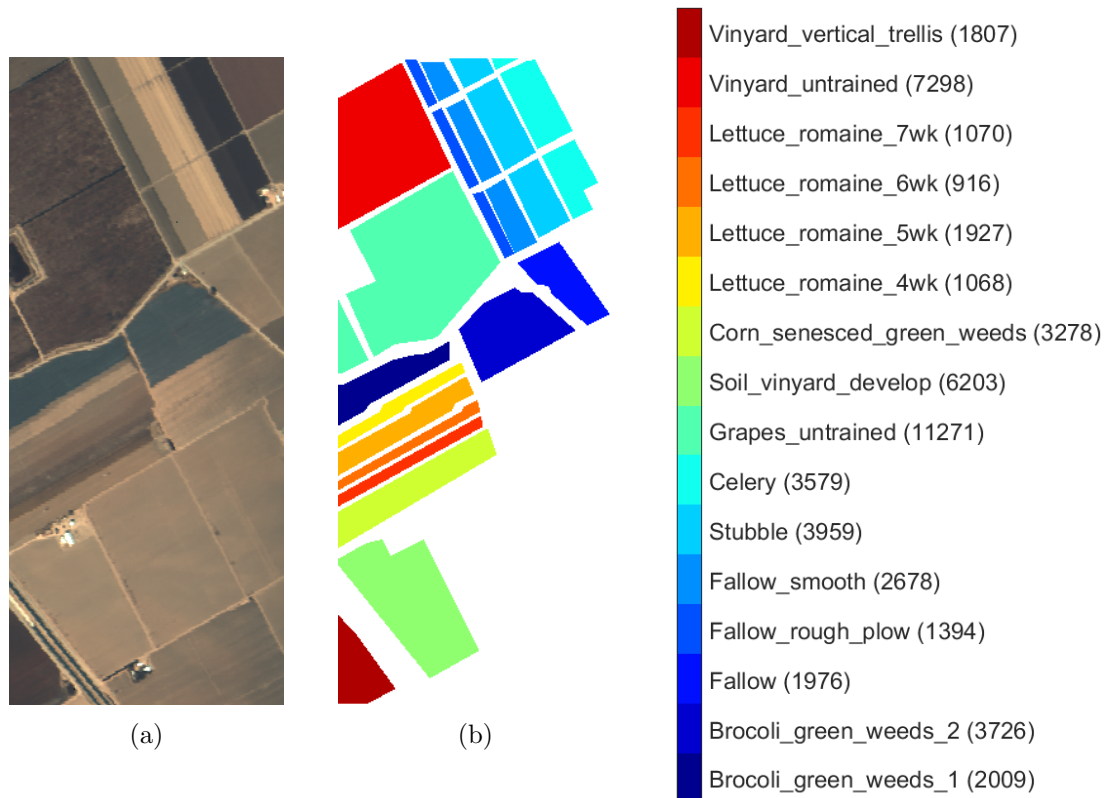
Figure A.2: Salinas dataset. (a) Pseudo RGB representation and (b) ground truth. The number of samples per class are indicated in brackets behind the name.
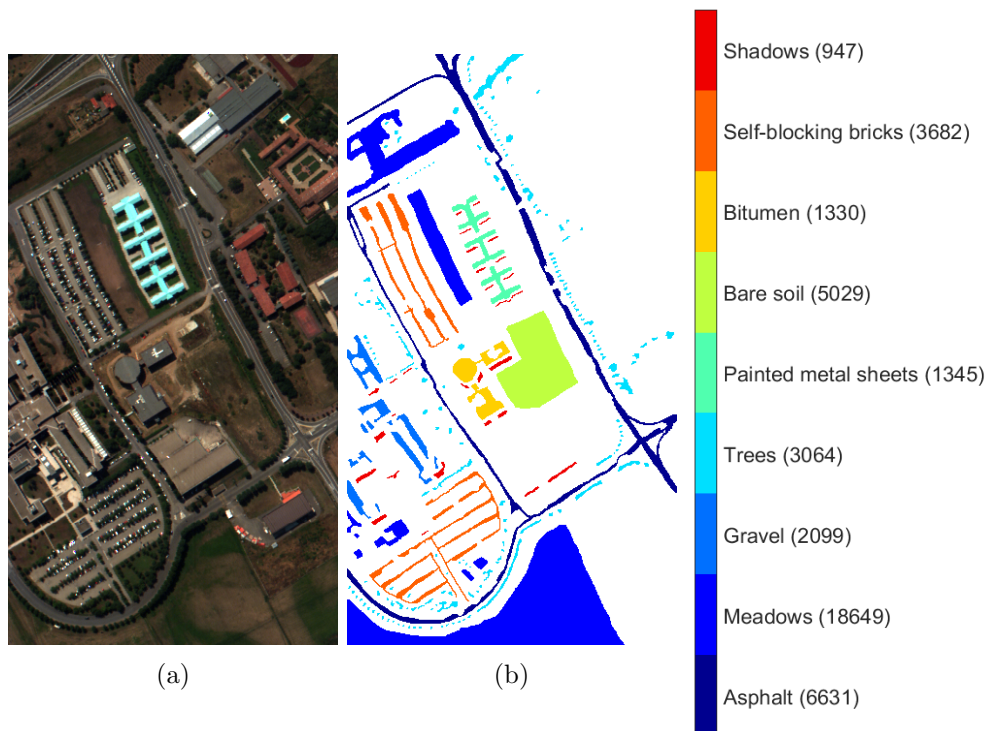
Figure A.3: Pavia University dataset. (a) Pseudo RGB representation and (b) ground truth. The number of samples per class are indicated in brackets behind the name.

Appendix B

List of Author's Publications

B.1 Journal publications

B.1.1 **Tschannerl, J.**, Ren, J., Zhao, H., Kao, F.-J. and Marshall, S., "Hyperspectral Image Reconstruction Using Multi-colour and Time-Multiplexed LED Illumination", *Optics and Lasers in Engineering*, vol. 121, pp 352-357, 2019

B.1.2 **Tschannerl, J.**, Ren, J., Yuen, P., Sun, G., Zhao, H., Yang, Z. and Marshall, S., "MIMR-DGSA: Unsupervised Hyperspectral Band Selection Based on Information Theory and a Modified Discrete Gravitational Search Algorithm", *Information Fusion*, vol. 51, pp 189-200, 2019

B.1.3 Zhao, J., Ren, J., Gao, J., **Tschannerl, J.**, Murray, P., Wang, D., "Automatic Events Extraction in Pre-stack Seismic Data Based on Edge Detection in Slant-stacked Peak Amplitude Profiles", *Journal of Petroleum Science and Engineering*, 2019

B.1.4 **Tschannerl, J.**, Ren, J., Jack, F., Krause, J., Zhao, H., Huang, W. and Marshall, S., "Potential of UV and SWIR hyperspectral imaging for determination of levels of phenolic flavour compounds in peated barley malt", *Food Chemistry*, vol. 270, pp. 105-112, 2018.

B.1.5 Mishra, P., Nordon, A., **Tschannerl, J.**, Lian, G., Redfern, S. and Marshall, S., "Near-infrared hyperspectral imaging for non-destructive classification of commercial tea products", *Journal of Food Engineering*, vol. 238, pp. 70-77, 2018.

B.2 Conference publications

B.2.1 **Tschannerl, J.**, Ren and J., Marshall, "Segmented Autoencoders for Unsupervised Embedded Hyperspectral Band Selection", in *7-th European Workshop on Visual Information Processing (EUVIP 2018)*, Tampere, Finland, 2018.

B.2.2 **Tschannerl, J.**, Ren, J. and Marshall, S., "Low Cost Hyperspectral Imaging Using Deep Learning Based Spectral Reconstruction", in *7th Hyperspectral Imaging and Applications Conference (HSI 2016)*, Coventry, United Kingdom, 2018.

B.2.3 **Tschannerl, J.**, Ren, J., Jack, F., Marshall, S. and Zhao, H., "Employing NIR-SWIR hyperspectral imaging to predict the smokiness of Scotch whisky", in *3rd Conference on Optical Characterization of Materials (OCM 2017)*, Karlsruhe, Germany, 2017.

B.2.4 Michael, K., Siti Salwa Binti Md Noor, S., **Tschannerl, J.**, Ren, J. and Marshall, S., in *Investigative Ophthalmology and Visual Science*, vol. 58, no. 8, Baltimore, MD, United States, 2017.

B.2.5 Mishra, P., Nordon, A., **Tschannerl, J.**, Lian, G., Redfern, S. and Marshall, S., "Near-infrared hyperspectral imaging for non-destructive classification of commercial tea products", in *18th International Conference on Near Infrared Spectroscopy (ICNIRS 2017)*, Copenhagen, Denmark, 2017.

B.2.6 **Tschannerl, J.**, Ren, J., McKee, D. and Marshall, S., "An integrated camera system for effective acquisition, capturing and transmission of hyperspectral data", in *6th Hyperspectral Imaging and Applications Conference (HSI 2016)*, Coventry, United Kingdom, 2016.

B.2.7 Noor, S. S. M., Michael, K., Marshall, S., Ren, **J., Tschannerl**, J. and Kao, F. J., "The properties of the cornea based on hyperspectral imaging: optical biomedical engineering perspective", in *23rd International Conference on Systems, Signals and Image Processing (IWSSIP 2016)*, Bratislava, Slovakia, 2016.