Automatisierte Analyse von Bauschuttrezyklaten auf der Basis von Bild- und Spektralinformationen

Dissertation
zur Erlangung des akademischen Grades

Doktoringenieur
(Dr.-Ing.)

vorgelegt der
Fakultät für Maschinenbau der
Technischen Universität Ilmenau

von Herrn
M.Sc. Petr Kuritcyn
geboren am 09. 04. 1989 in Leningrad/USSR

1. Gutachter: Univ.-Prof. Dr. rer. nat. Gunther Notni
2. Gutachter: apl. Prof. Dr.-Ing. habil. Ayoub Al-Hamadi
3. Gutachter: Prof. Dr.-Ing. habil. Thomas Ortlepp

Tag der Einreichung: 16.07.2019
Tag der wissenschaftlichen Aussprache: 20.08.2020

urn:nbn:de:gbv:ilm1-2020000692

Danksagung

Meine Dissertation entstand unter der Betreuung zweier Doktorväter – Univ.-Prof. Dr.-Ing. habil. Gerhard Linß und Univ.-Prof. Dr. rer. nat. Gunther Notni. Bei ihnen möchte ich mich besonders bedanken. Ohne Herrn Prof. Linß hätte ich die Arbeit nicht anfangen können. Daher bedanke ich mich für die gegebene Möglichkeit, die Promotion zu beginnen. Ich bin auch Herrn Prof. Notni sehr dankbar für die Übernahme der Leitung und die Unterstützung in der zweiten Hälfte meiner Promotion.

Meine Arbeit wurde im zweiten und dritten Jahr durch die Thüringer Graduiertenförderung finanziert. Ohne die Finanzierung wäre es für mich nicht möglich gewesen, an meiner Dissertation zu arbeiten. Deswegen möchte ich mich beim Freistaat Thüringen für die finanzielle Hilfe bedanken.

Durch eine Projektstelle als wissenschaftlicher Mitarbeiter konnte ich meine Arbeit abschließen. Die Projektförderung des Projektes RezykDetekt hat die Weiterarbeit und Beendigung meiner Dissertation ermöglicht. Dafür bin ich dem Bundesministerium für Wirtschaft und Energie dankbar.

Außerdem möchte ich mich bei allen Kollegen des Fachgebiets der Industriellen Bildverarbeitung und Qualitätssicherung bedanken, welche mich während der Arbeit mit Rat und Tat unterstützt haben. Mein besonderer Dank gilt Frau Dr.-Ing. habil. Katharina Anding für lange wissenschaftliche Diskussionen, die geleistete Unterstützung und die Zusammenarbeit bei mehreren gemeinsamen Publikationen.

Ich möchte mich auch besonders bei meiner Familie bedanken, die mich trotz eines großen Abstandes und mehrerer Staatsgrenzen während meines langen Studiums immer unterstützt und mir geholfen haben.

Abschließend bedanke ich mich herzlich bei meiner Partnerin Elena für die Unterstützung und die langen Korrekturnächte, in denen wir zusammen gearbeitet haben.

Kurzfassung

Bau- und Abbruchabfälle stellen Gemische aus mineralischen, metallischen und organischen Anteilen dar, welche eine entsprechend hochwertige Aufbereitung benötigen, um als rezyklierte Gesteinskörnungen wieder im Herstellungsprozess von Baustoffen verwendet werden zu können. Gesteinskörnungen stellen ein körniges Material dar, welches für die Betonherstellung geeignet ist. In Deutschland wie auch weltweit wird nur ein sehr geringer Teil des anfallenden Bauschutts für die Herstellung von Beton wiederverwendet. Variierende Gehalte an porösen Partikeln, Kontamination durch organisches und anorganisches Material erschweren die Sortierung der Gemische. Stand der Technik im Bereich der Analyse der Bau- und Abbruchabfälle ist die manuelle Inspektion durch Laborassistenten, weil nicht alle Bauschuttklassen zurzeit mittels automatisierter Methoden klassifiziert werden können. Für die Erkennung und Separation von aufbereiteten Bauabfällen wurden bisher nur einige gezielte Untersuchungen vorgenommen. Die Untersuchungen zeigten, dass nur ein modernes optisches System als Kombination von zwei oder mehreren spektralen sowie auch ortsaufgelösten Sensoren unter Verwendung adaptierter Erkennungsverfahren zukünftig in der Lage sein könnte, die Vielzahl der Stoffe im Bauschutt zuverlässig unterscheiden zu können. Die Automatisierung der Erkennung von Schüttgütern, insbesondere Bauschuttrezyklaten, würde zu einer enormen Zeit- und damit auch Kostenersparnis führen. Die Lösung einer komplexen Aufgabe wie die Bauschutterkennung benötigt die Anwendung verschiedener Algorithmen aus den Bereichen des maschinellen Lernens, der Bildverarbeitung und der Spektroskopie. Daraus folgt, dass diverse Untersuchungen zur Datensatzerstellung und -strukturierung, Merkmalsextraktion und Auswahl der geeigneten Merkmale mittels Merkmalsselektionsverfahren, Auswahl der Klassifikationsalgorithmen und Anpassung des Klassifikators zur Lösung der Aufgabe durchgeführt werden müssen. Im Rahmen dieser Arbeit wurde eine automatisierte Analyse von Bauschuttrezyklaten auf Basis von Bild- und Spektralinformationen realisiert. Ein Analyseverfahren für die Qualitätssicherung rezyklierter Gesteinskörnungen wurde entwickelt. Eine Basis für die Untersuchungen stellen durch Laborspezialisten vorbereitete und vorsortierte Proben dar. Die Proben wurden mittels eines Aufbaus mit hochauflösender 3-CCD- Kamera, einer Kombination von Auflicht- und Durchlichtbeleuchtung und mit einem NIR-Spektrometer aufgenommen. Daraus ergeben sich drei Datensätze auf der Basis von Bild-, Spektral- und Hybrid-Information (Kombination von beiden Informationen). Unterschiedliche Algorithmen für die Merkmalsselektion und Merkmalsextraktion wurden auf den Datensätzen untersucht und angepasst. Für die Lösung der Erkennungsaufgabe wurden diese Algorithmen zusammen mit verschiedenen Klassifikatoren aus den Bereichen der statistischen Klassifikatoren (Naive Bayes), der Entscheidungsbäume (Random Forest), der instanzbasierten Klassifiktoren (k-Nächste Nachbarn), der Support-Vektor-Maschinen und der künstlichen neuronalen Netze. Das Erkennungsproblem stellt eine komplexe Optimierungsaufgabe mit einer Vielzahl an Einflussfaktoren dar. Die Faktoren wurden in der Arbeit beschrieben und untersucht. Die Fusion von Bild- und Spektralinformation sowie eine passende Optimierung von beiden Informationsteilen erlaubt im Ergebnis eine Gesamterkennungsrate von 99,9% unter Anwendung des Klassifikators SVM mit polynomialem Kern.

Abstract

Construction and demolition waste (CDW) is a heterogenous mixture, which contains mineral, metal and organic components. These mixtures require special preparation before they can be used as recycled aggregates for concrete production. Only a small part of CDW used worldwide and also in Germany for this purpose due to contamination by organic and mineral components and components with high porosity. State of the Art in CDW-Sorting is a manual analysis by laboratory specialists because existing automated methods cannot classify all components of CDW. Several studies were done for analysis of CDW. They showed that a combination of multiple spectral and image sensors can provide enough information to distinguish different construction material classes from each other. Automatization of recognition of CDW will save time and resources in future. A solution of this complex task requires the usage of different algorithms of machine learning, image processing and spectroscopy. It means that diverse investigations of dataset creation and structure, feature extraction/selection, classifier selection and adaptation of these algorithms for the given problem should be done. Automated analysis of CDW based on image and spectral information was realized in this work. An analysis method for recycled aggregates of CDW was developed. Investigations are based on manually prepared and sorted samples by laboratory specialists. These samples were captured by a system with high-resolution 3CCD-camera, a combination of incident and transmitted light illumination and a NIR-spectrometer. It results in three datasets, which based on image, spectral and hybrid information (combination of both). Diverse algorithms for feature selection and extraction were tested and adapted on these datasets. These algorithms were used together with different classifiers: probabilistic classifiers (Naive Bayes), decision trees (Random Forest), instance-based classifiers (k-Nearest Neighbors), support vector machines and neural networks. The recognition problem is a complex optimization task with many influencing factors. These factors were investigated and described in this work. The fusion of image and spectral information allowed in combination with an appropriate optimization to reach a total recognition rate of 99.9 % by using a SVM classifier with polynomial kernel.

Inhaltsverzeichnis

1. Einleitung

In einer Zeit immer höherer Anforderungen im Bereich der Lebenswissenschaften, der Produktsicherheit sowie der Ressourcen- und damit Umweltschonung im Rohstoffsektor erlangen automatisierte innovative Analyse- und Diagnoseverfahren, insbesondere bei einem gleichzeitigen exponentiellen Fortschritt im Bereich der Rechentechnik, eine immer größere Bedeutung. So können Quantensprünge in verschiedenen Wissenschaftsdisziplinen, wie z. B. der Nahrungsmittelsicherheit, der Umweltmesstechnik sowie der Rohstoffaufbereitung durch neue leistungsfähige optische Analyseverfahren überhaupt erst möglich gemacht werden. Optische Verfahren besitzen gegenüber anderen Verfahren (z. B. taktil, akustisch etc.) den großen Vorteil nicht invasiv zu sein, d.h. mit dem Untersuchungsmedium nicht zu interagieren und damit das Analyseergebnis nicht zu verfälschen. Ein weiterer Hauptvorteil ist die Zeiteffizienz optischer Verfahren gegenüber chemischen Analysen. Die Verbesserung der Ressourceneffizienz erlaubt eine aktive Reduktion von Umweltbelastungen und dient damit direkt dem aktiven Klimaschutz. Die Ressourceneffizienz gilt es bei einer Vielzahl von Rohstoffaufbereitungen und Recyclingverfahren zu steigern. Dabei muss die Klassifikation der im jeweiligen Anwendungsgebiet vorkommenden Objektklassen umgesetzt werden, um entsprechende Handlungsrichtlinien zu ermöglichen und somit z. B. Sortierprozesse optimal zu steuern.

1.1. Aufgabenstellung und Motivation

Aktuell zeigt sich insbesondere die Erschließung neuer Rohstofflager bei gleichzeitiger Reduktion des Abfallaufkommens als eine wesentliche Aufgabe, um insbesondere umwelt-politischen und wirtschaftlichen Zielen gleichermaßen Rechnung tragen zu können. Insbesondere der Abbruch von Bauwerken und gebauter Infrastruktur stellt anthropogene Stofflager dar, welche bei knapper werdenden Primärrohstoffvorkommen und -verteuerung zunehmend für die Sekundärrohstoffgewinnung genutzt werden müssen.

In Deutschland wie auch weltweit wird nur ein sehr geringer Teil des anfallenden Bauschutts für die Herstellung von Beton wiederverwendet. Große Mengen des Bauschutts werden lediglich als Bettungsmaterial im Straßen- und Wegebau eingesetzt. Der Bundesverband Baustoffe - Steine und Erden e.V. gibt für Deutschland im Jahr 2014 insgesamt 202,0 Mio. Tonnen Bau- und Abbruchabfälle an (Abbildung 1).

Die aufbereiteten mineralischen Abbruchabfälle werden größtenteils (in 2014 zu 81,0 M.-%) im Erd- und Straßenbau als Hinterfüllungen und Tragschichten eingesetzt, da hierfür nur geringe technische Anforderungen gelten. Nur ein sehr geringer Anteil von 17,0 M.- % (meistens in der Asphaltherstellung) fließt wieder in den Hochbau zurück. Grund für den begrenzten Einsatz ist die hohe Heterogenität des aufbereiteten Bauschutts. Sekundärrohstofflager stellen Gemische aus mineralischen, metallischen und organischen Bestandteilen dar, welche nur nach einer entsprechenden Aufbereitung als rezyklierte Gesteinskörnungen wieder eingesetzt werden können. Rezyklierte Gesteinskörnungen werden im Rahmen der Qualitätssicherung und Güteüberwachung hinsichtlich der stofflichen Zusammensetzung charakterisiert. Anhand der Zusammensetzung werden die Verwertungsmöglichkeiten bzw. möglichen Einsatzgebiete (Liefertypen) definiert. Die Güteüberwachung ist die Voraussetzung zur Erlangung des Produktstatus. Die exakte Erkennung der unterschiedlichen Bestandteile ist nicht nur für die Qualitätssicherung von hoher Bedeutung, sondern stellt auch die Grundvoraussetzung für die Sortierung dar, um die Qualität der Rezyklate zu verbessern und die Störstofffracht in

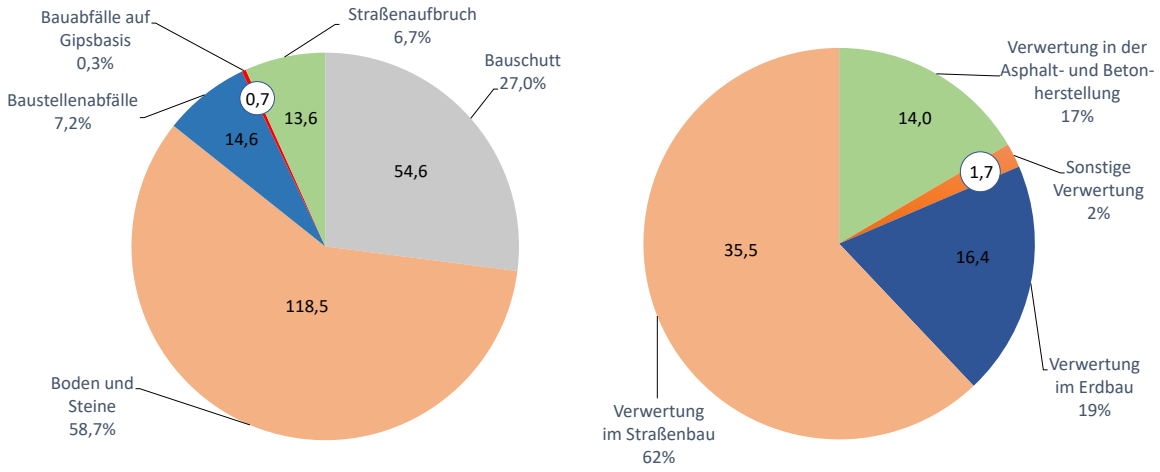

Abbildung 1: Statistisch erfasste Mengen mineralischer Bauabfälle 2014 (in Mio. t) und Verwertung der Recycling-Baustoffe 2014 (in Mio. t) [Kreislaufwirtschaft Bau, 2017]

den Recyclingbaustoffen (z. B. Recyclingbeton) zu reduzieren. Ein sortenreines Recycling würde eine Verwertung auf höherem Niveau und die Herstellung neuer Produkte ermöglichen. Dafür werden Verfahren einerseits für die Analytik und andererseits für die Stofftrennung der rezyklierten Gesteinskörnungen sowohl in die Hauptgruppen (z. B. Beton, Gesteinskörnung, Kalksandstein, Ziegel, Gips, Asphalt, Glas usw.) als auch innerhalb dieser Gruppen (z. B. Normal-, Leicht- und Porenbeton) benötigt. Aktuelle umweltpolitische und rechtliche Randbedingungen wirken sich zukünftig auf die Materialwirtschaft in der Beton- und Mauerstein- sowie Gipsindustrie deutlich verändernd aus. Zur Erzielung von zukunftsfähigen, effizienten und nachhaltigen Ressourcenströmen und Wertschöpfungsketten bedarf es daher innovativer Technologien und Konzepte.

Auf dem Recyclingsektor finden hauptsächlich traditionelle Sortiertechniken, welche mit dem Ansatz der mechanischen Trennung arbeiten, Anwendung. Diverse mechanische Prozesse, wie z. B. die magnetische Abscheidung von Eisenschrott, die Fraktionierung durch Siebung und die Abscheidung von Leichtstoffen durch Windsichter sowie die Setzmaschinentechnik, sind etabliert. Es gibt intensive Forschungen, um die traditionellen Sortiertechniken, die mit dem Ansatz der mechanischen Trennung arbeiten, abzulösen. Grund dafür ist einerseits der unzureichende Separationseffekt bei Stoffen sehr ähnlicher Dichte und Kornform und andererseits die stetige Zunahme der Produktvielfalt, welche die traditionellen Techniken an ihre Grenzen stoßen lässt. Seit einiger Zeit werden in bestimmten Sektoren der Recyclingwirtschaft (Papier, Kunststoff, Glas, Holz) sensorgestützte Sortierverfahren eingesetzt, um den Sortierprozess exakter, schneller und preisgünstiger zu gestalten. Hierbei handelt es sich allerdings weniger um Analysesysteme für die stoffliche Zusammensetzung, sondern vielmehr um Sortiersysteme geringer Klassenkomplexität und großen Massendurchsatzes. Bei diesen Geräten kommen hauptsächlich optische und magnetische Sensoren, Nah-Infrarot- und Röntgenstrahlsensoren zur Anwendung. Die Verfahren ermöglichen auf Basis von charakteristischen Eigenschaften, wie z.B. Farbe, Form oder Absorption bestimmter Wellenlängen, eine Unterscheidung verschiedener Proben. Für die Lösung einfacher Erkennungsaufgaben ist es ausreichend visuelle Eigenschaften, wie Farbe oder Form, aufzunehmen und zusammen

mit einfachen Algorithmen zu nutzen. Komplexe Aufgaben mit einer großen Klassenvielfalt, wie die Erkennung von Bauschuttrezyklaten, benötigen für die Lösung zum einen deutlich mehr Informationen, wie z.B. spezifische Oberflächenstrukturmerkmale oder zusätzlich chemische und physikalische Eigenschaften, und zum anderen komplexe Algorithmen aus dem Bereich des maschinellen Lernens. Die zusätzliche Information kann mittels hochauflösender Kameras, Spektrometer, Hyperspektral-Kameras u.a. erfasst werden. Für die Erkennung und Separation von aufbereiteten Bauabfällen wurden bisher nur einige gezielte Untersuchungen vorgenommen [Linss et al., 2012a], [Anding et al., 2011], [Linss et al., 2017], [Hollstein et al., 2017]. Die Untersuchungen zeigten, dass eine alleinige Anwendung eines Kamerasensors oder alternativ von spektralen Informationen für die Erkennung von Bauabfällen nicht ausreicht und nur ein modernes optisches System als Kombination von zwei oder mehreren spektralen sowie auch ortsaufgelösten Sensoren unter Verwendung adaptierter Erkennungsverfahren zukünftig in der Lage sein könnte, die Vielzahl der Stoffe im Bauschutt zuverlässig unterscheiden zu können. Die Automatisierung der Erkennung von Schüttgütern, insbesondere Bauschuttrezyklaten, würde zu einer enormen Zeit- und damit auch Kostenersparnis führen.

1.2. Ziele

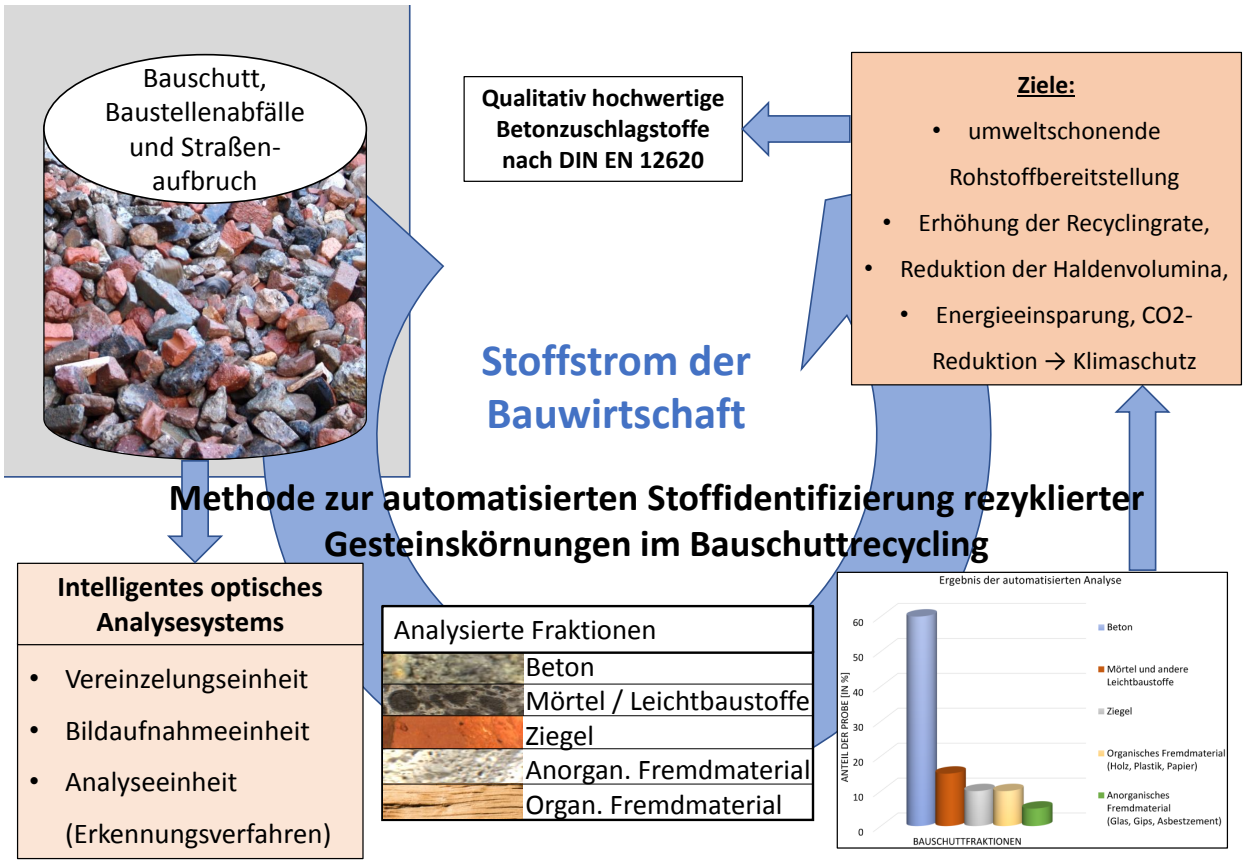

Abbildung 2: Prinzipdarstellung des intelligenten optischen Systems zur automatisierten Qualitätssicherung von Bauschuttrezyklaten mit lernfähiger Bildanalyse (in Anlehnung an [Anding, 2010])

Im Rahmen dieser Arbeit soll eine automatisierte Analyse von Bauschuttrezyklaten auf der Basis von Bild- und Spektralinformationen realisiert werden, um den Schlüsselfaktoren „Erhöhung der Zuverlässigkeit“ und „Steigerung der Ressourceneffizienz“ gerecht zu werden. Das Forschungsziel besteht darin, eine Analyseeinheit für die Qualitätssicherung rezyklierter Gesteinskörnungen, wie beispielsweise aufgeschlossenen Beton- und Mauerwerkbruch, zu entwickeln (siehe Abbildung 2). Dabei liegt der Fokus insbesondere auf Untersuchungen von der Anwendbarkeit unterschiedlicher Algorithmen aus dem Bereich maschinelles Lernen. Untersuchungen auf Bildern, Spektren und der daraus kombinierten Information müssen durchgeführt und miteinander verglichen werden. Hier gilt es insbesondere spezifische spektrale Eigenschaften des Probengutes zu untersuchen, um wellenlängenspezifische Charakteristiken optimal für die Erkennungsalgorithmen nutzen zu können. Zudem muss eine Vielzahl verschiedener Bildmerkmale ausgewählt und untersucht werden. Außerdem müssen Klassifikationsalgorithmen an die gegebene Aufgabe angepasst sowie eine adaptierte Erkennungsroutine entwickelt werden.

Auf Basis der für das Baustoffrecycling gegebenen Normen müssen Anforderungen an die zu erreichenden Erkennungsraten des Systems erarbeitet werden. Das zu entwickelnde System soll in der Lage sein, Bauschuttmaterialien mit einer hohen Sicherheit gemäß der gesetzlichen Standards zu erkennen.

1.3. Inhalt der Arbeit

Der erste Teil der Dissertation beschreibt den Stand der Technik zu spezifischen und ausgewählten Aspekten der gegebenen Erkennungsaufgabe. Die Bildverarbeitungsschritte und Schritte des Spektrenanalysesystems, welche ein Basis für die optische Bauschutterkennung darstellen, werden im Kapitel 2 beschrieben. Die grundlegenden Begriffe des maschinellen Lernens, der Spektroskopie und Bildverarbeitung werden wegen der häufigen Verwendung in der Arbeit als erstes erklärt. Dann werden einige Komponenten der Bildverarbeitungskette näher beschrieben. Hierbei liegt ein Hauptfokus auf Algorithmen für Merkmalsselektion/Merkmalsextraktion und Klassifikationsalgorithmen, welche in den praktischen Untersuchungen der Arbeit sehr oft angewendet werden. Ausgewählte Klassifikatoren und Merkmalsselektion-/Merkmalsextraktionalgorithmen werden explizit beschrieben. Im Kapitel 3 werden die Gesteinskörnungen charakterisiert und wichtige Standards im Bereich der Betonherstellung erklärt. Eine Liste von ähnlichen Gerätelösungen im Bereich der Bauschutterkennung wird im Kapitel 4 dargestellt und der Unterschied zwischen diesen und dem zu entwickelnden eigenen Verfahren erklärt. Im zweiten Teil der Arbeit werden Anforderungen an das Aufnahmesystem formuliert, sowie Vorüberlegungen zu geeigneten Merkmalen und Datensatzstrukturierung durchgeführt. Praktische Untersuchungen werden im dritten Teil der Arbeit beschrieben. Die Aufnahme der Spektren und Bilder wird in den Kapiteln 10 und 11 entsprechend dargestellt. Die Anwendung von ausgewählten Algorithmen wird in den Kapiteln 12, 13 und 14 ausführlich beschrieben. Die Kapitel umfassen praktische Untersuchungen auf drei Datensätzen: Bilddatensatz, Spektraldatensatz und Hybrid-Datensatz. Die Ergebnisse werden im Kapitel 14 verglichen. Die Zusammenfassung und ein Ausblick zu weiteren Arbeiten wird am Ende der Arbeit in den Kapiteln 15 und 16 gegeben.

Teil I.
Stand der Technik

2. Grundlagen für die automatisierte Bauschutterkennung

Die Lösung einer komplexen Aufgabe, wie der optischen Bauschutterkennung, erfordert die Anwendung diverser Verfahren aus den Bereichen der Bildverarbeitung, der Spektroskopie und des maschinellen Lernens. Die grundlegenden Begriffe und ausgewählte Verfahren aus diesen Bereichen, welche in der Arbeit Anwendung finden werden, sollen in diesem Kapitel näher beschrieben werden.

2.1. Bildverarbeitungschritte in der Objekterkennung

Der Zweck des Bildverarbeitungssystems für die Objekterkennung besteht darin, dass die Objekte auf der Basis visueller Eigenschaften der einen oder anderen Klasse zugeordnet werden können. Einige Schritte des Bildverarbeitungssystems für die Objekterkennung sind mit Algorithmen aus dem Bereich des maschinellen Lernens verbunden. Deswegen ist es notwendig, zuerst grundlegende Begriffe des maschinellen Lernens zu erklären.

Grundlegende Begriffe des maschinellen Lernens

Das maschinelle Lernen stellt einen Satz von Methoden dar, welche automatisch Muster in Daten erkennen können und nutzen die erhaltene Information, um neue Daten vorherzusagen oder andere Entscheidungen unter Unsicherheit zu machen [Murphy, 2012]. Maschinelles Lernen löst eine Vielzahl an gegebenen Problemen. Man unterscheidet zwischen [Mohri et al., 2012]:

- **Klassifikation**
  Einordnung der Input-Daten in gegebene Kategorien [Bishop, 2006]
- **Regression**
  Einordnung der Input-Daten zu einer reellen kontinuierlichen Variablen [Bishop, 2006].
- **Dimensionsreduktion**
  Umwandlung der initialen Darstellung der Objekte in eine einfachere; Beibehalten einiger Eigenschaften der initialen Informationen [Mohri et al., 2012].
- **Ranking**
  Einordnung der Objekte laut eines Kriteriums [Mohri et al., 2012].
- **Clustering**
  Erkennung von Gruppen ähnlicher Objekte in Daten [Bishop, 2006].

Ein **Exemplar** ist im Maschinellen Lernen eine Instanz oder Probe, die für Training bzw. Test genutzt wird [Mohri et al., 2012].

Ein **Merkmal** ist im Maschinellen Lernen ein Attribut eines Exemplares [Mohri et al., 2012]. Es enthält verschiedene Informationen über das Objekt und ist in Form von Text oder numerischen Werten gegeben.

Das **A-priori-Wissen** ist von vornherein gewonnene Kenntnisse, z.B. ein Expertenwert.

Ein **Merkmalsvektor** ist ein Merkmalssatz, welcher ein Objekt näher beschreibt.

Schritte des Bildverarbeitungssystems für die Objekterkennung

Die Lösung der Erkennungsaufgabe mittels eines Bildverarbeitungssystems wird in folgenden Schritte ausgeführt [Demant et al., 2011], [Gonzalez and Woods, 2007]:

- Bildaufnahme
- Segmentierung
- Merkmalsextraktion
- Merkmalsselektion
- Erkennung/Klassifikation

Jeder Schritt dieser Kette wird im Weiteren näher beschrieben.

2.2. Bildaufnahme

Digitale Bilder können unter Anwendung unterschiedlicher Sensoren erstellt werden [Demant et al., 2011]. Die Mehrheit der Bilder ist von der Kombination aus Beleuchtung und Reflexion oder Absorption der Objekte in der Szene generiert. Die Beleuchtung kann von unterschiedlichen Quellen kommen, wie typische elektromagnetische Quellen: Radar, Infrarot oder Röntgensystem, oder von nicht traditionellen Quellen, wie Ultraschall oder Beleuchtungsmuster von einem Computer. Abhängig von der Quelle kann die elektromagnetische Energie reflektiert oder transmittiert werden [Gonzalez and Woods, 2007].

Die Mehrheit der heutigen Sensoren basieren auf den Halbleitersensoren. Sie transformieren die Intensität der einfallenden elektromagnetischen Strahlung in eine Spannung bzw. in einen Digitalwert, welcher abhängig von der Anwendung weiter verwendet und verarbeitet werden kann. Man unterscheidet zwei Typen von Bildsensoren [Demant et al., 2011]:

- Anhand der geometrischen Anordnung von einzelnen Elementen (s.g. Pixel) - Zeilen- und Matrixsensoren
- Anhand der Funktionsweise und Fertigungstechnologie - Charge Coupled Device (abgekürzt CCD) und Complementary Metal Oxide Semiconductor (abgekürzt CMOS) Sensoren

Matrixsensoren bestehen aus mehreren Spalten und Zeilen und nehmen ein ganzes Bild während einer einzigen Belichtungszeit auf. Zeilensensoren im Gegensatz dazu bestehen nur aus einer Zeile und nehmen die Zeilen nacheinander auf, um ein Bild zu bekommen [Demant et al., 2011].

Farbbilder in der digitalen Bildverarbeitung müssen in einem Farbraum dargestellt werden, welcher eine Basis für die gesamte Bildverarbeitungskette ermöglicht. Es ist notwendig, einen geeigneten Farbraum auszuwählen, um nachfolgende Bildverarbeitungsalgorithmen erfolgreich anwenden zu können.

Farbraum als Basis für die Bildverarbeitung

RGB-Farbraum

Im RGB-Farbraum ist jede Farbe durch drei spektrale Komponenten - rot, grün und blau dargestellt. Das Modell basiert auf dem kartesischen Koordinatensystem. Der RGB-Farbraum ist auf der Abbildung 3 dargestellt.

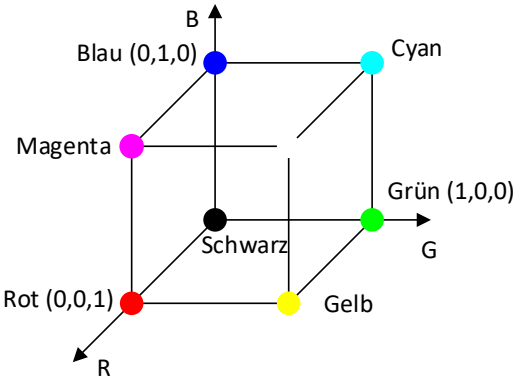

Abbildung 3: RGB-Farbraum

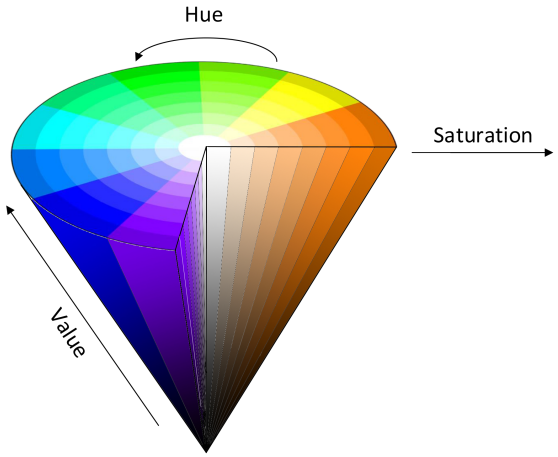

Abbildung 4: HSI-Farbraum

Drei Ecken stellen die Hauptfarben (rot, grün, blau) dar und die anderen drei - die sekundären Farben (gelb, cyan, magenta). Schwarz liegt am Anfang und Weiß befindet sich an der davon am weitesten entfernten Ecke. Die verschiedenen Farben im Modell sind die Punkte auf oder innerhalb des Kubus.Es wird davon ausgegangen, dass alle Farbwerte normiert sind, deswegen liegen alle Werte von R, G und B im Bereich von 0 bis 1 (Abbildung 3). Weil man in der Bildverarbeitung die nicht normierten Werte nutzt, liegen die im Bereich von 0 bis 255. Für das Graustufenbild entspricht 0 - Schwarz und 255 - Weiß [Gonzalez and Woods, 2007].

HSI-Farbraum

Der RGB-Farbraum ist ideal für die Hardware-Implementierung, aber dieses Modell ist schlecht geeignet für die Farbbeschreibung hinsichtlich der menschlichen Wahrnehmung und Interpretation. Dafür wurde das HSI (Hue, Saturation, Intensity)-Farbmodell entwickelt. In diesem Modell ist die Intensität von farbtragender Information (Hue und Saturation) entkoppelt. Im Ergebnis basiert der HSI-Farbraum auf den natürlichen und intuitiven Beschreibungen für Menschen, die letztendlich die Bildverarbeitungsalgorithmen entwickeln [Gonzalez and Woods, 2007].

Der HSI-Farbraum ist in der Abbildung 4 dargestellt. Die vertikale Achse repräsentiert die Intensität, der Farbton (Hue) hängt von dem Winkel zu einem Referenzpunkt ab. Nor-

malerweise entspricht der Winkel von 0° Rot und der Farbton vergrößert sich gegen den Uhrzeigersinn. Die Sättigung (Saturation) ist die Länge des Vektors vom Koordinatenursprung bis zum Punkt. Es ist wert zu erwähnen, dass der Koordinatenursprung sich bei der Kreuzung von Farbplane und vertikaler Achse befindet [Gonzalez and Woods, 2007].

Das Bild kann aus dem RGB-Farbraum in den HSI-Farbraum umgewandelt werden [Gonzalez and Woods, 2007]:

$$Hue = \begin{cases} \theta & \text{if} \quad Blue \leq Green \\ 360 - \beta & \text{if} \quad Blue > Green \end{cases} \tag{1}$$

wo β ist [Gonzalez and Woods, 2007]:

$$\beta = \arccos \left\{ \frac{\frac{1}{2}[(Red - Green) + (Red - Blue)]}{[(R - Green)^2 + (Red - Blue)(Green - Blue)]^{1/2}} \right\} \tag{2}$$

$$Saturation = 1 - \frac{3}{(Red + Green + Blue)}[\min(Red, Green, Blue)] \tag{3}$$

$$Intensity = \frac{1}{3}(Red + Green + Blue) \tag{4}$$

Es wird angenommen, dass die Farbkomponenten Red, Green und Blue normiert [0,1] sind.

Die Anwendung des HSI-Farbraums hat sich für verschiedene Erkennungsaufgaben bewährt. Er ist einer der am meisten verwendeten Farbräume in der Bildverarbeitung [Anding, 2010], [Anding et al., 2011], [Süße and Rodner, 2014], [Burger and Burge, 2015]. Deswegen wird der Farbraum ebenfalls für die Untersuchungen in dieser Arbeit verwendet und die aufgenommenen RGB-Bilder in HSI-Bilder mittels der Formeln 1, 2, 3, 4 transformiert.

Beleuchtung

Die Beleuchtung ist in der Regel individuell für jede konkrete Aufgabe und soll experimentell angepasst werden. Die oben erwähnten Beleuchtungsquellen können abhängig vom Problem unterschiedlich angeordnet werden. Man unterscheidet zwei Beleuchtungsanordnungen: Auflicht und Durchlicht. Die Beleuchtung und die Kamera befinden sich auf der gleichen Seite des Objekts beim Auflicht und auf zwei verschiedenen Seiten beim Durchlicht. Im Rahmen der Anordnungen können die Lichtquellen unterschiedliche Positionen und Formen haben, sowie der Winkel des Lichteinfalls kann variieren. So lassen sich unterschiedliche Effekte erzielen. Die Auflichtbeleuchtung kann folgende Formen haben [Demant et al., 2011]:

- **Diffuses Auflicht** kann starke Reflexionen und Schatten reduzieren

- **Gerichtetes Auflicht**: Hellfeldbeleuchtung für die Kontrolle von Bohrungen und Dunkelfeldbeleuchtung für die Verbesserung des Kontrasts von Oberflächenrauheiten und Vertiefungen bzw. Erhöhungen
- **Polarisiertes Licht** wird zur Vermeidung der Reflexionen verwendet
- **Ringbeleuchtung** kann intensiv und schattenfrei die Objekten beleuchten
- **Beleuchtung im Strahlengang** für die Kontrolle von Innenbohrungen und für die Anwendung in der Endoskopie
- **Strukturierte Beleuchtung** zur Feststellung der dreidimensionalen Eigenschaften unter geringem Helligkeitskontrast

Nähere Beschreibung der Methoden findet man in [Gonzalez and Woods, 2007], [Demant et al., 2011].

Für die Bildaufnahme in der Arbeit wurde der bereits entwickelte Aufbau verwendet. Die nähere Beschreibung der Komponenten findet man in Kapitel 10.

2.3. Segmentierung

Während der Bildaufnahme werden oft nicht nur Objekte aufgenommen, sondern auch andere Elemente der Bildszene, wie z.B. Hintergrund. Deswegen ist es notwendig festzustellen, welche Pixel des Bildes zum Objekt gehören und das Objekt vom Hintergrund für eine weitere Bildanalyse zu isolieren. Diese Operation heißt Segmentierung. Zurzeit existieren viele Segmentierungsverfahren. Die wichtigsten und am weitesten verbreiteten sind die Binärsegmentierung, die Konturverfolgung, die Kantendetektion und die intelligente Segmentierung auf Basis von Bildmerkmalen mittels Klassifikatoren. Diese Methoden werden näher in [Jähne, 2012], [Demant et al., 2011], [Gonzalez and Woods, 2007] beschrieben.

2.4. Merkmalsextraktion aus dem Farbbild

In diesem Kapitel wird ein Überblick über die wesentlichsten und meist verwendeten Merkmalsalgorithmen in der Bildverarbeitung gegeben.

2.4.1. Konturmerkmale aus dem Farbbild

Aus der Vielzahl möglicher, aus dem Farbbild berechenbarer Konturmerkmale (siehe auch [Demant et al., 2011]) sollen hier beispielhaft nur einige ausgewählte Merkmale genannt werden.

Flächeninhalt

Die Fläche des Objektes ist die Anzahl der Pixel, die innerhalb der Objektkontur liegen [Demant et al., 2011].

Schwerpunkt

Die Schwerpunktskoordinaten (x_s, y_s) des Objektes können wie folgt berechnet werden [Demant et al., 2011]:

$$x_s = \frac{\sum_i x_k}{A} \qquad y_s = \frac{\sum_i y_k}{A} \tag{5}$$

wo A Flächeninhalt des Objektes ist, x_k und y_k die Koordinaten des Objektes sind

Rechteckförmigkeit

Das Merkmal Rechteckförmigkeit stellt ein Verhältnis der Objektfläche A zur Fläche A_{kur} des kleinsten umschreibenden Rechtecks dar [Demant et al., 2011]:

$$R = \frac{A}{A_{kur}} \tag{6}$$

Radius
Der Radius charakterisiert einen Abstand zwischen dem Schwerpunkt des Objektes und irgendeinem Punkt seiner Kontur. Man unterscheidet minimale, mittlere und maximale Radien.

2.4.2. Farbmerkmale aus dem Farbbild

Aus der Grauwertsverteilung des Objektes können verschiedene Farbmerkmale berechnet werden [Abmayr, 1994, Demant et al., 2011]:

- minimaler Grauwert
- maximaler Grauwert
- Standardabweichung der Grauwerte
- mittlere Informationsinhalt (Entropie) usw.

2.4.3. Texturmerkmale aus dem Farbbild

Die Textur stellt die Oberflächenstruktur dar. Es gibt verschiedene Methoden, um die Texturen quantitativ zu beschreiben. Viele dieser Methoden basieren auf den statistischen Einsätzen (Statistik 2. Ordnung). Einer der einfachsten Algorithmen basiert auf dem Gradient, Grauwertübergänge zwischen benachbarten Pixeln [Demant et al., 2011].

In [Laws, 1980] wurde die Methode vorgeschlagen, einen Filter für die Transformation des Bildes zu nutzen und dann die Texturenergie anhand der absoluten Summe der Pixel in der Nachbarschaft von jedem Pixel zu berechnen. Die Filter sind auf der Basis von 5 Masken aufgebaut: L - Level, E - Edge, S - Spot, R - Ripple und W - Wave, alle stellen eindimensionale Vektoren dar. Die Multiplikation der Masken ergibt einen Filter. Insgesamt existieren 16 Laws-Filter, die alle für die quantitative Texturbeschreibung genutzt werden können.

2.5. Spektrenanalyse

Eine andere Informationsquelle für die optische Objekterkennung stellt die Spektralinformation dar. Die Methoden der Spektrenanalyse ermöglichen die Objekterkennung auf der Basis charakteristischer spektraler Eigenschaften, welche jedes Material hat.

- **Spektrenanalyse**
  Mehrzahl experimenteller Methoden, welche die Absorption, Emission und Streuung von elektromagnetische Strahlung auf Molekülen und Atomen betreffen [Hollas, 2004].

- **Spektrum** (in der Spektrenanalyse)
  ein Umfang absorbierter, abgestrahlter oder gestreuter elektromagnetischen Strahlung von Molekülen und Atomen der Materie [Daintith, 2008].

Das elektromagnetische Spektrum stellt die Gesamtheit aller Wellenlängen von Gammastrahlen bis hin zu Radiowellen dar. Auf der Basis der Wellenlänge bzw. Frequenz wurde das gesamte elektromagnetische Spektrum in folgende Bereiche eingeteilt: Gamma-Strahlung, Röntgenstrahlung, Ultraviolette Strahlung, sichtbare Strahlung (Licht), Infrarot, Mikrowellen und Radiowellen [Bernath, 2005]. Die Wechselwirkung der Strahlung mit der Materie sind das Sachgebiet der Wissenschaft der Spektroskopie [Skoog et al., 2013]. Man unterscheidet folgende Prozesse: Transmission, Streuung, Reflexion, Absorption und Emission [Sabins, 2007]. Das größte Interesse stellen jedoch die Prozesse dar, bei denen der Elektronenübergang zwischen zwei verschiedenen Atomniveaus auftritt, d.h. Absorption und Emission. Andere Wechselwirkungen sind mehr von Schüttguteigenschaften des Materials als von spezifischen Eigenschaften der Atome und Moleküle abhängig [Skoog et al., 2013]. Es existieren mehrere spektroskopische Methoden. Die oben genannten Prozesse ermöglichen die Lösung verschiedener Aufgaben im Rahmen der qualitativen bzw. quantitativen Analyse. Aus allen Wechselwirkungen der Strahlung mit Materie ist die Absorption am häufigsten verwendbar für alle Spektralbereiche [Hollas, 2004].

Schritte des Spektrenanalysesystems für die Objekterkennung

Der Zweck des Spektrenanalysesystems für die Objekterkennung besteht darin, dass die Objekten auf der Basis von charakteristischen spektralen Eigenschaften der einen oder anderen Klasse zugeordnet werden können. Die Lösung der Aufgabe wird in ähnlichen Schritte ausgeführt wie bei der Bildverarbeitung (siehe Kapitel 2.1), ausgenommen die ersten zwei Schritte (markiert mit *-Zeichen):

- *Spektrenaufnahme
- *Vorverarbeitung des Spektrums
- Merkmalsextraktion
- Merkmalsselektion
- Erkennung/Klassifikation

Die zwei neuen Schritte dieser Kette werden im Weiteren näher beschrieben.

2.5.1. Spektrenaufnahme

Das allgemeine Schema für die Spektrenaufnahme ist in Abb. 5 dargestellt.

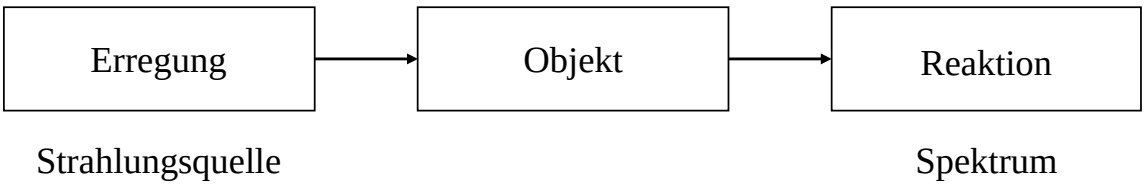

Abbildung 5: Allgemeine Struktur der Spektrenaufnahme (in Anlehnung an [Skoog et al., 2006])

Als Anregung können Licht, Wärme, Strom, Partikel oder chemische Reaktionen verwendet werden.

Die Reaktion wird als die ausgestrahlte Energie beim Übergang des physikalischen Systems von einem angeregten Niveau zum Grundzustand oder als die Menge der absorbierten Energie zur Erregung des physikalischen Systems gemessen [Skoog et al., 2013].

Instrumente für die Spektroskopie enthalten die fünf nachfolgenden Komponenten: eine stabile Strahlungsenergiequelle, einen Wellenlängen-Selektor, einen Probenbehälter, den Strahlendetektor sowie ein Signalverarbeitungs- und Anzeigegerät [Skoog et al., 2013]. Für Absorptionsmessungen werden externe Strahlungsquellen genutzt. Diese müssen ausreichend stark und stabil sein, um eine einfache Detektion bzw. Messungen zu gewährleisten. Es gibt zwei Typen von Strahlungsquellen: 1) mit kontinuierlichem Spektrum, die eine Intensitätsverteilung ohne große Schwankungen in einem Wellenlängenbereich haben, und 2) Linienquellen, die nur bestimmte Wellenlänge ausstrahlen und jede von denen umfasst einen kleinen benachbarten Wellenlängenbereich [Skoog et al., 2013]. Die Quellen werden zudem in kontinuierliche und impulsartige unterteilt. Als kontinuierliche Strahlungsquellen werden für die optische Spektroskopie Gasentladungslampen im Ultraviolett (UV), Xenonbogenlampen im UV/sichtbaren Bereich (VIS), Halogen-Glühlampen und Nernstlampen im VIS/Infrarot (IR), Nickelchrom-Draht und Globar im IR genutzt [Skoog et al., 2013]. Die Anwendungshäufigkeit der Halogen-Glühlampen ist immer höher wegen des breiten Wellenlängenbereichs, der langen Haltbarkeit und einer guten Lichtintensität. Außer den kontinuierlichen Quellen finden auch Linienquellen, wie Quecksilberdampflampen im UV/VIS und Laser für bestimmte Spektroskopiearten Anwendung.

Der Wellenlängen-Selektor begrenzt die Strahlung auf wenige Wellenlängen, die aufgenommen und gemessen werden können. Das verbessert die Empfindlichkeit und die Selektivität. Als Selektoren können Monochromatoren oder Filter für die Selektion des engen Wellenlängenbereiches angewendet werden. Für die Verteilung und Aufnahme des breiteren Spektralbereiches wird der Spektrograph genutzt. Im Prinzip haben alle Selektoren außer Filtern ein dispergierendes Element, z.B. ein optisches Gitter oder ein Prisma, welches das einfallende Licht stufenlos auffächert. Im Monochromator wird mittels des Austrittsspalts nur ein enger Wellenlängenbereich isoliert und hindurchgelassen. Im Spektrograph nimmt

dagegen ein Multi-Wellenlängen-Sensor das ausgebreitete Licht auf. Filter lassen nur begrenzte Wellenlängen durch und reflektieren bzw. absorbieren die restliche Strahlung.

Ein Detektor ist ein Gerät, das ermittelt, misst oder die Änderungen einer Variablen in der Umgebung anzeigt. In modernen Instrumenten wird die Information in Form des elektrischen Signals umgewandelt und verarbeitet. Der Transducer verwandelt die anfallende Strahlung in die elektrischen Parameter Spannung, Strom oder in elektrische Ladung. Es existieren zwei Typen der Strahlungs-Transducer: einer reagiert auf Photonen, ein anderer auf Wärme. Alle Photonen-Detektoren basieren auf der Wechselwirkung der Strahlung mit der Oberfläche des Sensors, welche Elektronen abstrahlt (Photoemission) oder den elektrischen Strom leiten kann (Photoleitfähigkeit). Photonenleiter können im gesamten IR-Bereich arbeiten. Im Prinzip misst man im IR die Temperatur des Sensors, der die Strahlung des Objektes aufnimmt oder die Änderung in der Leitfähigkeit als Folge von Absorption der einfallenden Strahlung. Der zu messende Temperaturunterschied ist gering und die Umgebungstemperatur kann das Ergebnis stark beeinflussen, was oft die Messgenauigkeit für Detektoren im IR-Bereich begrenzt. [Skoog et al., 2013]

Der Signalprozessor ist ein elektronisches Gerät, welches das elektrische Signal vom Detektor verstärken und konvertieren und auch die Phase ändern sowie unerwünschte Komponenten löschen kann [Skoog et al., 2013]. Meistens werden Computer verwendet, da sie sehr funktional sind und unterschiedliche Aufgaben lösen können, wie z.B.: Daten speichern, austauschen mit anderen Systemen, bearbeiten usw.

Die Kombination von obengenannten Komponenten ergibt zwei Typen von Instrumenten für Absorptionsmessungen. Das Spektrometer ist ein Gerät, das einen Mono- oder Polychromator für die Wellenlängenselektion nutzt, und die Strahlung in ein elektrisches Signal verwandelt. Das Spektralphotometer ist ein Spektrometer, welches das Verhältnis zwischen zwei Strahlen, abstrahlende und ankommende, ausmessen kann. Das Photometer hat im Gegensatz dazu einen Filter, der nur einen engen Wellenlängenbereich durchlässt. Photometer haben einen Vorteil im Preis und in der Einfachheit des Aufbaus, während Spektrometer und Spektralphotometer kontinuierliche Spektren aufnehmen können.

Das Ergebnis des Instrumenteneinsatzes wird in Form eines Spektrums, welches die Abhängigkeit der Absorption bzw. der Emission von der Wellenlänge (Wellenzahl, Frequenz) darstellt, aufgenommen und weiter analysiert. Im Prinzip hat jede Substanz eine einmalige spektrale Abtastung [Skoog et al., 2013] und verschiedene Wellenlängenbereiche haben diverse Mechanismen der Wechselwirkung mit Materie und liefern unterschiedliche Informationen: im VIS-Bereich finden die Änderungen der Elektronenverteilung statt, Gamma-Strahlung nimmt Einfluss auf die Kernkonfiguration, IR-Strahlung weist chemisch funktionale Gruppen in der Substanz nach, usw. Verschiedene Qualitätssicherungsaufgaben benötigen unterschiedliche Methoden, die auf der richtigen Auswahl des Spektralbereiches basieren.

Die Infrarot-Spektroskopie ist eine der häufigsten Methoden für die Untersuchungen organischer und anorganischer Stoffe. Im Prinzip sind es Absorptionsmessungen der Proben auf dem Strahlengang, welche unter Anwendung verschiedener Komponenten für die Untersuchung von Gasen, Flüssigkeiten und Stoffen angewendet werden können. Der wichtigste Vorteil dieser Methode ist die Identifikation unterschiedlicher Funktionsgruppen, was sehr hilfreich für die Erkennung organischer bzw. anorganischer Stoffe ist [Settle, 1997].

Alle Messergebnisse enthalten zwei Teile: das Signal, welches die relevante Information enthält, und das Rauschen, welches nicht informativen oder sinnlosen Inhalt trägt. Spektroskopische Analysen sind mit zwei Typen des Rauschens beeinträchtigt, des chemischen

und des instrumentell bedingten Rauschens. Das chemische Rauschen kommt aus einer nicht messbaren Änderungen in den Umgebungsparametern, wie z.B.: Schwankungen des Luftdruckes, der Temperatur und der Luftfeuchtigkeit, welche das chemische Gleichgewicht beeinflussen oder Vibrationen ändern die Verteilung der Proben. Das instrumentelle Rauschen betrifft alle Komponenten des Messsystems, wie den Detektor, die Strahlungsquelle, die Signalverarbeitungsgeräte usw. Bei mehreren Messungen ist die Rauschstärke konstant und unabhängig von der Signalstärke. Dadurch wird der Einfluss des Rauschens auf das Ergebnis immer höher mit Verringerung des Signals. Deshalb ist das Verhältnis des Signals S zum Rauschen N besser nutzbar für die Charakterisierung der Qualität der Analysemethode [Skoog et al., 2006].

Ein Schwerpunkt bei der Spektrenaufnahme ist die Verbesserung des Signal-Rausch-Verhältnisses, welches sehr stark auf das Ergebnis und die Genauigkeit der Messungen Einfluss nimmt. Eine verbreitete Methode ist der Ensemblemittelwert (*ensemble averaging*), bei dem mehrere Spektren aufgenommen und gemittelt werden. Es existieren verschiedene Methoden für die Eliminierung irregulärer Schwankungen im Signal, was zur Verbesserung des Signal-Rausch-Verhältnisses führt: die Segmentmittelung (*boxcar averaging*), die Kleinste-Quadrate-Approximation, die Polynomglättung, der gleitende Mittelwert (*moving average*), der exponentiell geglättete Mittelwert (*exponential moving average*) u.a. Diese Methoden werden in [Skoog et al., 2006], [Skoog et al., 2013], [Savitzky and Golay, 1964] näher beschrieben.

2.5.2. Chemometrische Merkmale aus dem Spektrum und Vorverarbeitung des Spektrums

Spektroskopische Methoden generieren große Datenvolumen in geringer Zeit. Alle Anwendungen im Bereich des maschinelles Lernens brauchen signifikante Informationen, um richtige Ergebnisse zu produzieren. Dafür ist es wichtig zu entscheiden, welche Daten aus einem großen Datenvolumen verwendet werden können und welche ausgeschlossen werden müssen. Manchmal ist es notwendig, eine manuelle oder automatisierte *subset selection* oder multivariate Datenanalyse anzuwenden, um dies zu erreichen [Baudelet, 2014]. Als Ergebnis erhält man eine reduzierte Datenmenge mit hohem Informationsgehalt.

Die einfachste Anwendung ist die Nutzung des gesamten Merkmalssatzes ohne gezielte Merkmalsauswahl, d.h. im Fall des Spektrums, die Verwendung aller aufgenommener Wellenlängen. Damit braucht man keine Entscheidung zu treffen, welche Daten für eine weitere Analyse zu nutzen sind. Des Weiteren besteht kein Risiko des Informationsverlustes. Der Nachteil ist allerdings ein höherer Rechenaufwand. Maschinelle Lernverfahren leiden häufig unter dem sogenannten Fluch der Dimensionalität (*curse of dimensionality*), welcher aussagt, dass die Anforderungen an die Größe der Datenmenge exponentiell mit dem Anstieg der Dimensionalität der Daten steigen. Um dieses Problem zu vermeiden, enthalten mehrere Lernverfahren verschiedene Dimensionalitätsreduktionstechniken [Baudelet, 2014].

Die Anwendung des Erfahrungswissens eines Experten im Bereich des maschinellen Lernens bringt oft die besten Ergebnisse. Die Auswahl eines geeigneten Merkmalssatzes, welcher auf der Bewertung des Experten basiert, berücksichtigt den physikalischen bzw. chemischen Hintergrund der Aufgabe, was zu einer optimalen Lösung mit geringerem Aufwand führt [Baudelet, 2014]. In der Spektroskopie beschränkt sich die Expertenwahl hinsichtlich der Merkmalsextraktion/Merkmalsselektion auf die am besten geeigneten Wellenlängen, welche

die charakteristischen Absorption- oder Emissionsbande für die relevanten Stoffe enthalten.

Obwohl die Anwendung des Expertenwissens häufig bereits eine gute Effizienz zeigt, kommt es je nach beeinflussenden Umgebungseffekten zu schlechteren Ergebnissen, so können z.B. die Proben eine komplexe Zusammensetzung haben, was zu schlechteren Erkennungsleistungen führt und eine Expertenanalyse schwieriger macht. In diesem Fall stellen automatisierte Merkmalsselektionsverfahren eine gute Alternative zur Expertenauswahl der Merkmale dar. Für die Auswahl der am besten geeigneten Merkmale wird oft eine Bewertungsfunktion oder eine spezielle Metrik verwendet, welche in einem Ranking höhere Funktionswerte für Objekteigenschaften mit höherer Signifikanz ausgibt [Baudelet, 2014]. Die Berechnung der Metrik für jeden Teildatensatz hilft bei der Suche des geeignetsten Merkmalssatzes. Leider vergrößert die umfangreiche Suche nach der Auswahl der n geeigneten Merkmale aus dem m dimensionalen Merkmalsdatensatz drastisch die Anzahl der notwendigen Berechnungen und bei hochdimensionalen Datensätzen wird eine solche Suche in einer begrenzten Zeit unausführbar. Deshalb finden in der Spektroskopie oft Approximationsverfahren Anwendung.

Im Endeffekt stellen alle Wellenlängen einen Merkmalssatz dar, der weiter komplett oder teilweise (z.B. nur charakteristische Absorptionsbande) im Lernprozess genutzt wird.

Die Daten werden durch verschiedene Umgebungsfaktoren über die Zeit beeinflusst, was zu Änderungen der Ergebnisse führt. Manche Vorverarbeitungsmethoden können auf Daten angewendet werden, bevor diese weiter im Lernprozess genutzt werden.

Einen der größten Einflüsse auf die Ergebnisse hat die Kopplung des Sensors an die Proben, was zu großen Schwankungen in der absoluten Intensität führt. Um das zu vermeiden, kann man die Daten I_i bezüglich ihres Mittelwertes $\bar{I}$ zentrieren (Gleichung 7) und normalisieren [Gasteiger and Engel, 2003].

$$I_{i(zentriert)} = I_i - \bar{I} = I_i - \frac{\sum_{i=1}^{n} I_i}{n} \tag{7}$$

wo n Anzahl der Messungen ist.

Die erste Methode eliminiert die absoluten Schwankungen in den Daten, die zweite verringert den Einfluss von relativen Änderungen.

Die Normalisierungsmethoden können mit wenig Aufwand die Robustheit des Systems erhöhen.

Eine andere Alternative zur Normalisierung ist die Verwendung der Ableitung der Spektraldaten [Siesler et al., 2002]:

$$\frac{dI}{d\lambda} = \frac{I_{i+1} - I_i}{\Delta\lambda} \tag{8}$$

λ stellt die Wellenlänge dar.

Die Ableitung wurde zur Entfernung des Hintergrundsignals oder zur Verbesserung der Auflösung verwendet [Siesler et al., 2002]. Die Gegenwart von kleinen Peaks kann besser im abgeleiteten Spektrum nachgewiesen werden [Adams, 2004]. Der Nachteil besteht darin, dass das abgeleitete Spektrum anfälliger für Rauschen ist. Das begrenzt den Anwendungsbereich

auf UV-, VIS- und NIR-Spektroskopie, wo die notwendige Anforderung zum Signal-Rausch-Verhältnis erfüllt werden kann, obwohl bei der Verringerung der Signalstärke auch der Informationsgehalt reduziert wird, d.h. die Ableitung des Spektrums hat oft zu wenig Einfluss auf das Signal-Rausch-Verhältnis [Siesler et al., 2002].

2.6. Merkmalsextraktionsverfahren

Eines der größten Probleme im maschinellen Lernen ist der Fluch der Dimensionalität: die Aufgaben, die gut im kleinen Merkmalsraum lösbar sind, sind oft unlösbar im mehrdimensionalen [Duda et al., 2001]. Um dieses Phänomen zu umgehen, werden verschiedene Methoden der Dimensionsreduktion angewendet.

Hauptkomponentenanalyse

Die Hauptkomponentenanalyse (englisch - *Principal Component Analysis (PCA)*), auch bekannt als Karhunen-Loève-Transformation, ist eine der meist verwendeten Methoden für die Dimensionsreduktion, Datenkompression, Merkmalsextraktion und Datenvisualisierung [Jolliffe, 2002]. Die *PCA* ist im Prinzip eine Projektion der Daten von einem hochdimensionalen in einen niedrigdimensionalen Merkmalsraum. Die Hauptkomponentenanalyse umfasst die Verwandlung und Drehung der originalen Achsen, welche die originalen Variablen darstellen, in neue Hauptachsen [Adams, 2004]. Jedes Objekt kann als ein Datenpunkt im mehrdimensionalen Raum beschrieben werden. Die Daten stellen eine Punktewolke dar. In dieser Wolke wird eine Achse derart gefunden, dass in deren Richtung die größte Varianz der Daten repräsentiert ist. Das ist die erste Hauptkomponente (*first principal component (PC1)*) (Abbildung 6).

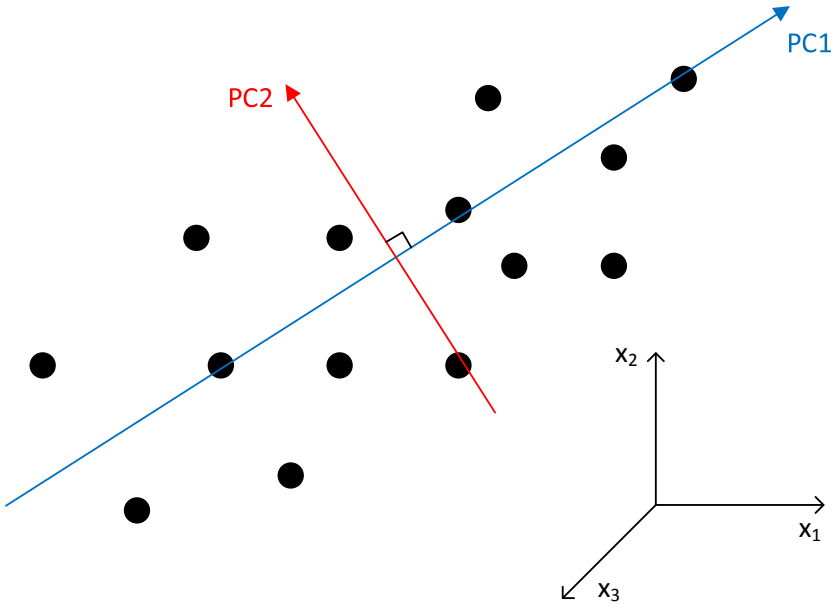

Abbildung 6: Berechnung der ersten zwei Hauptkomponenten (in Anlehnung an [Gasteiger and Engel, 2003])

Die zweite Komponente (*PC2*) ist orthogonal zur ersten *PC* und stellt die zweitgrößte Varianz dar. Weitere Komponenten werden entsprechend berechnet [Gasteiger and Engel, 2003]. Weil die neuen Achsen orthogonal zueinander sind, sind die neuen Variablen unkorreliert [Adams, 2004]. Das führt zu einer Verringerung der Freiheitsgrade der Daten und als Folge dessen zu einem reduzierten Rechenaufwand [Duda et al., 2001]. Ein anderes Anwendungsgebiet ist die Datenvorverarbeitung. In diesem Fall wird der Datensatz umgewandelt, um die Merkmale zu normalisieren, dies spielt insbesondere für Erkennungsaufgaben und die Mustererkennung eine bedeutende Rolle [Bishop, 2006].

Diskriminanzanalyse

Die Idee hinter dieser Methode ist die Bestimmung eines niedrigdimensionalen Merkmalsraums, in welchem die d-dimensionalen Daten trennbar sind. Die Trennbarkeit wird als statistische Werte in Form der Varianz und des Mittelwertes gemessen [Xanthopoulos et al., 2013].

Die Diskriminanzfunktion ist eine Funktion, die den Eingangsvektor $\mathbf{x}$ verwendet und ihm eine von K Klassen, bezeichnet C_k, zuordnet. Die einfachste Diskriminanzfunktion ist [Bishop, 2006]:

$$y(\mathrm{x}) = \mathrm{w}^{\mathrm{T}}\mathrm{x} + b \tag{9}$$

wo w der Gewichtsvektor und b der Schwellwert ist.

Geometrisch bedeutet das, wenn $\|\mathrm{w}\| = 1$ ist, sind alle y_i Projektionen auf einer Linie in Richtung $\mathbf{w}$ [Duda et al., 2001]. Der Eingangsvektor wird bei $y(\mathrm{x}) \geq 0$ der Klasse c_1 zugeordnet, ansonsten der Klasse c_2. Die Grenze wird als $y(\mathrm{x}) = 0$ bestimmt. Es ist erwünscht, dass die Projektion eine bestmögliche Trennung zwischen den Klassen erlaubt. Dabei wird deutlich, dass, wenn die originalen Daten sehr starke Überlappungen zwischen den Klassen aufweisen, sogar die besten Gewichtsvektoren $\mathbf{w}$ keine ausreichende Trennung ermöglichen, sodass diese Methode an ihre Grenzen stößt . Die Trennbarkeit wird in Form des Mittelwertes wie folgt berechnet [Duda et al., 2001]:

$$\mathrm{m_i} = \frac{1}{n_i} \sum_{\mathrm{x} \in c_i} \mathrm{x} \tag{10}$$

wo n_i die Anzahl der Datenpunkte der Klasse c_i darstellt.

Dann berechnet sich die Projektion des Mittelwertes wie folgt [Duda et al., 2001]:

$$\tilde{m}_i = \frac{1}{n_i} \sum_{y \in c_i} y \tag{11}$$

Unter Berücksichtigung der Gleichung 9, erhält man [Duda et al., 2001]:

$$\tilde{m}_i = \frac{1}{n_i} \sum_{y \in c_i} \mathrm{w}^{\mathrm{T}}\mathrm{x} = \mathrm{w}^{\mathrm{T}}\mathrm{m_i} \tag{12}$$

Dies führt zur Distanzberechnung zwischen den Klassen nach [Duda et al., 2001]:

$$|\tilde{m}_1 - \tilde{m}_2| = |\mathrm{w}^{\mathrm{T}}(\mathrm{m}_1 - \mathrm{m}_2)| \tag{13}$$

Das bedeutet, dass die Distanz zwischen den Klassen mit der Vergrößerung **w** skaliert werden kann. Außer dem maximalen Abstand zwischen den Klassen ist eine kleinstmögliche Varianz innerhalb der Klassen wichtig. Die Varianz für die projizierten Daten wird nachfolgend berechnet [Duda et al., 2001]:

$$\tilde{s}_i = \sum_{y \in c_i} (y - \tilde{m}_i)^2 \tag{14}$$

Die Summe $\tilde{s}_1^2$ und $\tilde{s}_2^2$ entspricht der gemeinsamen Intraklassenvarianz (*within-class scatter*). Sowohl diesen Wert als auch die Interklassendistanz enthält das nachfolgend beschriebene Kriterium J [Duda et al., 2001]:

$$J(\mathrm{w}) = \frac{|\mathrm{m}_1 - \mathrm{m}_2|^2}{\tilde{s}_1^2 + \tilde{s}_2^2} \tag{15}$$

Dieses Kriterium maximiert die Diskriminanzfunktion. Der Wert **w** führt während der Maximierung des Kriteriums J auch zur Minimierung der Intraklassenvarianz und Maximierung der Interklassendistanz. Um **w** zu bestimmen, werden die Varianzmatrizen (*scatter matrices*) definiert [Duda et al., 2001]:

$$S_i = \sum_{x \in c_i} (x - m_i)(x - m_i)^T \tag{16}$$

$$S_W = S_1 + S_2 \tag{17}$$

Unter Berücksichtigung der Gleichungen 9, 14 und 16 erhält man [Duda et al., 2001]:

$$\tilde{s}_i^2 = \sum_{x \in c_i} \mathrm{w}^T (x - m_i)(x - m_i)^T \mathrm{w} = \mathrm{w}^T S_i \mathrm{w} \tag{18}$$

sowie die Summe dieser Varianzen [Duda et al., 2001]:

$$\tilde{s}_1^2 + \tilde{s}_2^2 = \mathrm{w}^T S_W \mathrm{w} \tag{19}$$

Ebenso wird die Interklassendistanz wie folgt berechnet [Duda et al., 2001]:

$$(\tilde{m}_1 - \tilde{m}_2)^2 = \mathrm{w}^T(m_1 - m_2)(m_1 - m_2)^T\mathrm{w} = \mathrm{w}^T S_B \mathrm{w} \tag{20}$$

mit $S_B = (m_1 - m_2)(m_1 - m_2)^T$.

S_W ist die Intraklassenvarianzmatrix (*within-class scatter matrix*) und S_B die Interklassendistanzmatrix (*between-class distance matrix*). In Form von S_W und S_B wird das Kriterium J wie folgt bestimmt [Duda et al., 2001]:

$$J(\mathrm{w}) = \frac{\mathrm{w}^T S_B \mathrm{w}}{\mathrm{w}^T S_W \mathrm{w}} \tag{21}$$

Die Gleichung 21 ist auch als Rayleigh-Koeffizient (*Rayleigh quotient*) bekannt. Die Formel für die Optimierung von J anhand von $\mathbf{w}$ ist wie folgt bestimmt [Duda et al., 2001]:

$$\mathrm{w} = \mathrm{S}_\mathrm{W}^{-1}(\mathrm{m}_1 - \mathrm{m}_2) \tag{22}$$

Die Lösung dieser Gleichung ergibt $\mathbf{w}$, bei dem das maximale Verhältnis der Interklassendistanz zur Intraklassenvarianz erreicht wird [Duda et al., 2001].

Die Datenprojektion auf die Hyperebene im Rahmen der Diskriminanzanalyse ist in Abbildung 7 dargestellt.

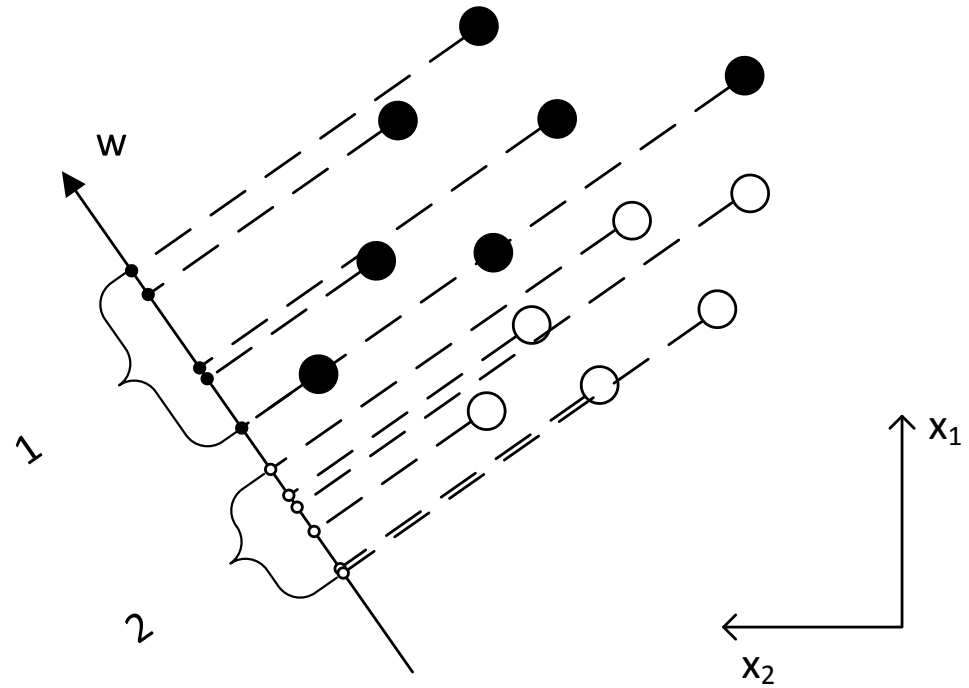

Abbildung 7: Datenprojektion auf die Hyperebene (in Anlehnung an [Xanthopoulos et al., 2013])

Die Diskriminanzanalyse kann für die Identifikation relevanter Merkmale, für die Dimensionsreduktion und auch für die Klassifikation unbekannter Proben genutzt werden [Xanthopoulos et al., 2013].

Der prinzipielle Unterschied zwischen der Diskriminanzanalyse und der Hauptkomponentenanalyse besteht darin, dass die Hauptkomponentenanalyse mit unmarkierten Daten arbeitet und die Varianz maximiert, während die Diskriminanzanalyse mit markierten Daten arbeitet und versucht die Trennung der Klassen zu maximieren [Zaki and Meira Jr., 2014].

2.7. Merkmalsselektionsverfahren

In der realen Welt enthalten die Daten im Gegensatz zur Theorie irrelevante oder kontraproduktive Information. Diese Information verringert die Leistung bei der Lösung der Aufgaben [Witten and Frank, 2011]. Wegen diesem negativen Einfluss werden außer den bereits beschriebenen Extraktions- und Dimensionreduktionsverfahren sogenannte Selektionsverfahren für die Eliminierung der irrelevanten Information angewendet. Der beste Weg für die Selektion ist die manuelle Auswahl, die auf einem guten Verständnis von Lernaufgabe und Bedeutung der Merkmale basiert. Wie im Kapitel 2.5.2 beschrieben, ist das jedoch nicht immer möglich und daher ist es notwendig automatisierte Verfahren anzuwenden.

Obwohl die Merkmalsselektion oft für die Auswahl der relevanten Information und die Eliminierung der irrelevanten genutzt wird, kann sie auch andere Anwendungsziele haben, wie z.B. [Guyon and Elisseeff, 2006]:

- Datenreduktion, um Speicheranforderungen zu begrenzen und die Rechengeschwindigkeit zu erhöhen;
- Verringerung des Merkmalssatzes für die Zeitersparnis bei zukünftigen Aufnahmen und bei der Vorverarbeitung;
- Leistungsverbesserung, um die Vorhersagegenauigkeit zu erhöhen;
- besseres Verständnis der Daten, um Kenntnisse über den Hintergrund der Probleme zu bekommen oder um die Daten zu visualisieren.

Die Merkmalsselektionsverfahren können in zwei Obergruppen unterschieden werden. Die erste Gruppe bewertet Merkmale auf der Basis allgemeiner Daten-eigenschaften, die zweite Gruppe führt die Bewertung eines Merkmalsteilsatzes (*feature subset*) mittels ausgewählter Methoden des maschinellen Lernens durch [Witten and Frank, 2011].

Die erste Gruppe arbeitet unabhängig vom Klassifikationsprozess, sie filtert irrelevante Merkmale heraus. Diese Verfahren sind sogenannte Filter-Verfahren. Diese Methoden basieren auf mathematischen Bewertungskriterien, welche direkt auf den Datensatz angewendet werden ohne eine Rückmeldung vom Klassifikator zu benötigen. Diese Verfahren produzieren im Gegensatz zu den beiden ersten Gruppen den geringsten Rechenaufwand [Guyon and Elisseeff, 2006].

Die zweite Gruppe kann in zwei Untergruppen geteilt werden. Die erste Untergruppe stellt die Methode dar, die direkt in den Klassifikationsprozess implementiert ist. Sie werden als sogenannte embedded-Methoden bezeichnet und sind fest mit entsprechenden Klassifikatoren verbunden. Die zweite Untergruppe enthält Verfahren, die rund um den Klassifikationsprozess

organisiert sind. Sie liefern dem Klassifikator einen Merkmalsteilsatz, um die Bewertung in Form gemessener Klassifikationsleistungen zu erhalten. Solche Verfahren werden als Wrapper-Methoden bezeichnet. [Guyon and Elisseeff, 2006].

Filterverfahren

Filtermethoden ergeben oft eine Liste von Merkmalen, die aufgrund der Relevanz in einer bestimmten Reihenfolge eingeordnet sind. Methoden für die Berechnung der Relevanz verwenden verschiedene statistische Kriterien, wie z. B. Korrelationskoeffizienten, welche die Abhängigkeitsstufe der Variablen von der Ausgabe bewerten, oder der klassische T-Test, F-Test und die Chi-squared-Statistik usw. Im Prinzip gehören solche Methoden zu den Filtern, welche die Merkmalsselektion ohne Anwendung der Klassifikatoren oder Prädiktoren durchführen [Guyon and Elisseeff, 2006].

Die Merkmale können entweder univariat oder multivariat bewertet werden [Aggarwal, 2014]. Univariate Filter bewerten das Merkmal unabhängig von anderen Merkmalen im Merkmalssatz. Multivariate Filter berücksichtigen im Gegensatz dazu den Zusammenhang und das Zusammenspiel der Merkmale im Merkmalssatz. Beide haben ihre Vor- und Nachteile.

Einer der verbreitetsten Einsätze der univariaten Methoden ist das Ranking der Merkmale aufgrund ihrer Relevanz. Das ist eine einfache und schnelle Methode, oft effektiv bei einer großen Anzahl an Merkmalen und kleinen Objektanzahlen. Dabei haben Merkmale, die eine bessere Trennung der Klassen ergeben, einen höheren Rang.

Aus der Vielzahl möglicher Filterverfahren sollen hier beispielhaft einige ausgewählte näher aufgeführt werden.

Korrelationskoeffizient

Der Pearson-Korrelationskoeffizient ist ein klassischer Wert zur Berechnung einer individuellen Merkmalsselektion. Der Korrelationskoeffizient laut Pearson berechnet sich wie folgt [Guyon and Elisseeff, 2006, Duda et al., 2001]:

$$\rho = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}} = \frac{\mathrm{Cov}(xy)}{\sigma_x \sigma_y} \tag{23}$$

wobei σ_x und σ_y den Quadratwurzeln der Varianz der Variablen x und y entsprechen und $\mathrm{Cov}(xy)$ der Kovarianz zwischen x und y.

Der Koeffizient kann Werte zwischen $\rho = -1$ und $\rho = 1$ annehmen. Bei der Analyse chemischer Daten soll die Korrelationsanalyse als erster Schritt angewendet werden. Falls zwei Merkmale hohe Korrelationskoeffizienten haben (z.B. $|\rho| > 0.9$), muss einer der redundanten Merkmale herausgefiltert werden [Gasteiger and Engel, 2003].

Information gain

Die informationstheoretische Maße werden häufig zur Merkmalsselektion verwendet. Eine von ihnen, die Entropie, ist ein Maß für den Informationsgehalt (die Information ist das Negativ der Entropie [Guyon and Elisseeff, 2006]) in der Datenverteilung [Duda et al., 2001]:

$$H(X) = -\sum_{i=1}^{n} P_i \log_2(P_i) \tag{24}$$

wobei P_i die Wahrscheinlichkeit für die zufällige Auswahl einer Probe ist, die zur Klasse i gehört ($i = 1..n$) und X ein Merkmal aus dem Merkmalssatz ist.

Der Wert der klassenbasierten Entropie liegt im Bereich von 0 bis $log_2(n)$. Eine höhere Entropie bedeutet eine stärkere Mischung der Klassen und ein Wert von 0 entspricht einer idealen Trennung und einer bestmöglichen diskriminativen Kraft des Merkmals [Aggarwal, 2015].

Mit der Entropie ist die sogenannte Transinformation, bekannt auch als gegenseitige Information (*mutual information*), Synentropie oder Informationsgewinn (*information gain*), eng verbunden. Sie ist ein Maß der Abhängigkeit zweier Variablen voneinander und zeigt die Verringerung der Unsicherheit über den Wert einer Variable aufgrund der Kenntnisse über eine andere Variable. Der Informationsgewinn wird zwischen Merkmal X und Klasse Y wie folgt berechnet [Aggarwal, 2014], [Duda et al., 2001]:

$$IG(X;Y) = H(X) - H(X|Y) = \sum_{x,y} g(x,y) \log \frac{g(x,y)}{p(x)p(y)} \tag{25}$$

wobei $p(x)$ und $p(y)$ den Verteilungen der Variablen x und y entsprechen und $g(x,y)$ die multivariate Verteilung x und y darstellt.

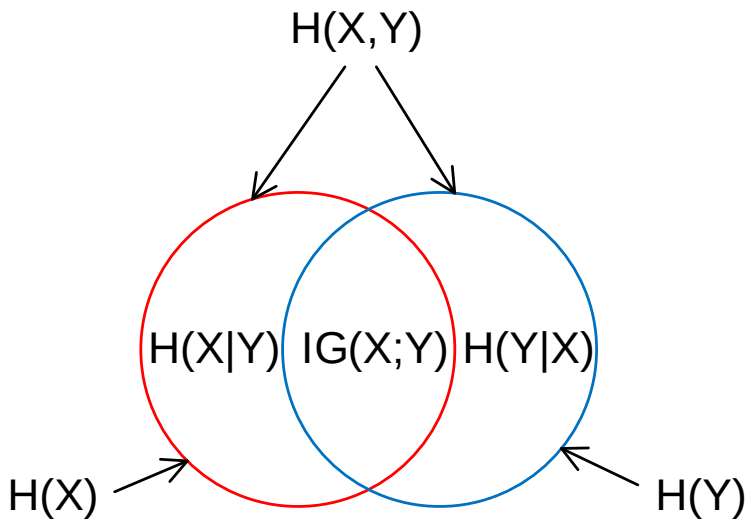

Abbildung 8: Mathematische Verhältnisse zwischen den Entropien $H(X)$ und $H(Y)$, dem Informationsgewinn $IG(X;Y)$, der bedingte Entropien $H(X|Y)$ und $H(Y|X)$ und der Kreuzentropie $H(X,Y)$ (in Anlehnung an [Duda et al., 2001])

Die mathematischen Verhältnisse zwischen Entropie und Informationsgewinn sind in Abbildung 8 dargestellt.

Die Idee hinter dem Informationsgewinn besteht darin, dass die Entropie vor und nach der Anwendung des Merkmals verglichen wird, so dass der Informationsgewinn bestimmt wird. Das Merkmal ist wichtiger, wenn der Informationsgewinn $IG(X;Y)$ zwischen der Zielverteilung (Klassenverteilung) und der Merkmalsverteilung größer ist [Guyon and Elisseeff, 2006].

Wegen der Effizienz der Rechenleistung und einfachen Interpretation ist der Informationsgewinn eine der häufigsten Methoden für die Merkmalsselektion. Es handelt sich um einen univariaten-Filter, deshalb können keine redundanten Merkmale gefunden und abgefiltert werden [Aggarwal, 2014].

χ^2-Statistik

Die Bestimmung der Korrelationskoeffizienten ist eine der einfachsten Methoden zur Bewertung von Merkmalen. Sie vermeidet Probleme mit der Beurteilung der Wahrscheinlichkeitsdichte, welche die Methode auf Basis der Informationstheorie hat [Guyon and Elisseeff, 2006]. Der Nachteil der Anwendung des linearen Korrelationskoeffizienten besteht darin, dass damit nur eine lineare Abhängigkeit zwischen Merkmalen bestimmt werden kann. Beim nichtlinearen Zusammenhang kann der Korrelationskoeffizient keine Abhängigkeit zwischen Merkmalen aufzeigen. Der Chi-Quadrat-Unabhängigkeitstest kann allerdings zuverlässig solche Abhängigkeiten bewerten [Runkler, 2010].

In der Statistik bilden die absoluten oder relativen Wahrscheinlichkeiten / Häufigkeiten von Variablen-Kombinationen eine Kreuztabelle. Die Tabelle kann analysiert werden, um die Korrelation zwischen den Variablen zu berechnen. Die Stärke des Zusammenhangs kann mittels Chi-Quadrat-Statistik bestimmt und wie folgt berechnet werden [Aggarwal, 2015]:

$$\chi^2 = \sum_{i=1}^{2^{|X|}} \frac{(O_i - E_i)^2}{E_i} \tag{26}$$

wobei O_i die reale Anzahl der Beobachtungen vom Typ i und E_i die erwartete Anzahl der Beobachtungen vom Typ i unter Annahme, dass zwei Variablen unabhängig voneinander sind, ist.

Offensichtlich ist χ^2 nur dann positiv oder Null, wenn alle realen und erwarteten Werte genau aufeinander abgestimmt sind. Je höher χ^2 ist, desto weniger stark ist die Unabhängigkeit der Variablen voneinander, sodass der Unterschied statistisch signifikant bei höherem χ^2 ist, so dass es möglich ist, die Hypothese über die Unabhängigkeit abzulehnen und das Merkmal als informativ einzuschätzen. Die Tabelle ergibt die Grenzwerte von χ^2 für verschiedene Signifikanzniveaus, welche die Möglichkeit geben, die Hypothese über die Unabhängigkeit der Variablen abzulehnen [Duda et al., 2001].

Chi-Quadrat Ergebnisse hängen nicht nur von der multivariaten Verteilung $g(x, y)$ ab, sondern auch von der Anzahl der Objekte n. Die Idee dahinter besteht darin, dass die Einschätzung der Wahrscheinlichkeit auf Basis der kleinen Auswahl nicht genau genug ist, deshalb ist die Signifikanz einer kleinen Korrelation eher gering [Guyon and Elisseeff, 2006].

ReliefF

Die Idee hinter dem Relief-Algorithmus besteht darin, dass der Wert des Merkmales auf der Fähigkeit zur Trennung der naheliegenden Instanzen basiert. Dafür sucht Relief für die zufällig ausgewählte Instanz x zwei Nachbarn: ein erster aus der gleichen Klasse (*nearest hit*, nächster Treffer $nearHit$) und ein zweiter aus der anderen Klasse (*nearest miss*, nächste Verfehlung $nearMiss$). Das ändert die Qualitätsbeurteilung w$[X]$ (Gewichtsvektor) für alle Merkmale X abhängig von ihren Werten für x, $nearMiss$ und $nearHit$ [Guyon and Elisseeff, 2006, Robnik-Šikonja and Kononenko, 2003]:

$$\mathrm{w}[X] = \mathrm{w}[X] - \frac{(|x - nearMiss| + |x - nearHit|)}{n \cdot (max(X) - min(X))} \tag{27}$$

wo n die Anzahl der Wiederholungen ist.

Wenn die Instanzen x und $nearHit$ unterschiedliche Werte X aufweisen, dann trennt das Merkmal X die zwei Instanzen aus einer Klasse, was nicht erwünscht ist. Deshalb verringert der Algorithmus die Qualitätsbeurteilung w$[X]$. Anderseits, wenn die Instanzen x und $nearMiss$ unterschiedliche Werte X haben, dann trennt das Merkmal X zwei unterschiedliche Klassen gut, was erwünscht ist, sodass Relief die Qualitätsbeurteilung w$[X]$ erhöht. Der ganze Prozess wiederholt sich n Mal, wobei n ein manuell einzustellender Parameter ist [Robnik-Šikonja and Kononenko, 2003].

Der ReliefF-Algorithmus ist eine Weiterentwicklung des originalen Relief-Algorithmus für den Multiklassen-Fall, welcher auch robuster ist und mit unvollständigen Daten arbeiten kann. ReliefF wählt, wie auch der Relief-Algorithmus, die Instanz x zufällig, sucht jedoch die k-nächsten Nachbarn aus der gleichen Klasse (*nearest hits*, nächste Treffer $nearHit_j$) sowie die k-nächsten Nachbarn aus jeder anderen Klasse (*nearest misses*, nächste Verfehlungen $nearMiss_j(c)$). Abhängig von den Werten des Merkmales X für die Instanzen $nearHit_j$ und $nearMiss_j(c)$ ändert sich die Qualitätsbeurteilung w$[X]$ für alle Merkmale X. Die Formel für die Umrechnung ist ähnlich zum Relief, mit der Ausnahme, dass die Beiträge von allen Treffern und Verfehlungen gemittelt werden. Der Beitrag für jede Klasse wird mit Zustimmung der Wahrscheinlichkeit der Klasse $P(c)$ berechnet (wird aus dem Training-Datensatz berechnet) [Robnik-Šikonja and Kononenko, 2003]:

$$\begin{aligned} \mathrm{w}[X] = \mathrm{w}[X] - \frac{1}{n \cdot k} \sum_{j=1}^{k} \frac{|x - nearHit_j|}{max(X) - min(X)} + \\ \frac{1}{n \cdot k} \sum_{c \neq class(X)} \left[\frac{P(c)}{1 - P(class(X))} \sum_{j=1}^{k} \frac{|x - nearMiss_j(c)|}{max(X) - min(X)} \right] \end{aligned} \tag{28}$$

Die Qualitätsbeurteilung w kann auch negativ sein, obwohl w$[X] \leq 0$ bedeutet, dass das Merkmal X irrelevant ist [Liu and Motoda, 2008].

Der Relief-Algorithmus stellt eine Methode zur Merkmalsselektion dar, welche nicht auf der Wahrscheinlichkeitsverteilung basiert [Guyon and Elisseeff, 2006]. Unter Berücksichtigung der Ähnlichkeit der Instanzen werden alle Merkmale implizit betrachtet [Liu and Motoda, 2008].

Der Einsatz der Filter-Verfahren ist nicht universell und hat folgende Nachteile, welche sich aus der Annahme einer Unabhängigkeit der Merkmale voneinander ergeben [Guyon and Elisseeff, 2006]:

- Merkmale, die allein irrelevant sind, können bei einer Kombination durchaus relevant sein;
- Merkmale, die relevant sind, können wegen einer Redundanz zu anderen Merkmalen trotz allem nutzlos sein.

Diese Nachteile treten bei der Anwendung von sogenannten multivariaten Filterverfahren nicht auf. Multivariate Methoden können damit potentiell eine bessere Leistung erreichen [Guyon and Elisseeff, 2006]. Eins der meist verwendeten Verfahren ist die *Correlation-based Feature Selection*.

Correlation-based Feature Selection

Wenn eine Gruppe von k Merkmalen gewählt wurde, kann man die Korrelationskoeffizienten zwischen den Merkmalen und Klassen berechnen, einschließlich auch der Kreuzkorrelation für die Merkmale. Der Wert der Merkmale nimmt mit Vergrößerung des Korrelationskoeffizienten zu und mit Erhöhung der Kreuzkorrelation ab. Angenommen wird, dass der mittlere Korrelationskoeffizient zwischen den Merkmalen und Klassen als $r_{ky} = \bar{\rho}(X_k, Y)$ bestimmt wird und der mittlere Kreuzkorrelationskoeffizient als $r_{kk} = \bar{\rho}(X_k, X_k)$ geschrieben wird. Der Koeffizient der Relevanz des Teildatensatzes kann dann wie folgt berechnet werden [Guyon and Elisseeff, 2006]:

$$J_M(X_k, Y) = \frac{k r_{ky}}{\sqrt{k + (k-1) r_{kk}}} \tag{29}$$

Diese Formel stellt den Pearson-Korrelationskoeffizienten mit allen normierten Variablen dar. Er wird mit Hinzufügen (*forward selection*) oder Löschen (*backward selection*) der Merkmale nacheinander im *Correlation-based Feature Selection(CFS)*-Algorithmus angewendet [Guyon and Elisseeff, 2006]. *CFS* ist ein einfaches Filter-Verfahren, welches die Merkmalsteilsätze entsprechend der Korrelationsfunktion (Gleichung 29) einordnet. Irrelevante Merkmale werden herausgefiltert, weil sie eine geringe Korrelation mit der Klasse haben. Redundante Merkmale werden herausgefiltert, wenn sie eine hohe Korrelation mit den restlichen Merkmalen aufweisen [Hall, 1998].

Der optimale Merkmalsteilsatz kann mittels drei verschiedener heuristischer Suchstrategien gefunden werden: *forward selection*, *backward elimination* und *best first*. *Forward selection* beginnt ohne Merkmale und fügt stetig (gierig - *greedy*) nacheinander weitere Merkmale hinzu, solange bis eine weitere Verbesserung nicht mehr möglich ist. *Backward elimination* beginnt mit dem ganzen Merkmalssatz und verringert die Anzahl der Merkmale nacheinander solange, bis eine Verschlechterung auftritt. *Best first* beginnt ohne Merkmale oder aber mit dem ganzen Merkmalssatz. Im ersten Fall fügt der Suchprozess einzelne Merkmale hinzu,

im zweiten entfernt er einzelne Merkmale. Um die Suche auf dem ganzen Merkmalssatz zu vermeiden, wird ein Kriterium für den Abbruch implementiert. Die Suche wird nach fünf aufeinanderfolgenden Merkmalsteilsätzen ohne Verbesserung gegenüber dem aktuellen abgebrochen [Hall, 1998].

CFS ist ein Filter und deshalb weniger rechenintensiv und unabhängig von den gewählten Lernalgorithmen. Es braucht $m((n^2-n)/2)$ Operationen für die Berechnung der Korrelationsmatrix für die Merkmale, wobei m die Anzahl der Objekte und n die Anzahl der Merkmale ist. Die Suche nach dem besten Merkmalssatz braucht höchstens $(n^2-n)/2$ Operationen für die *forward selection* oder *backward elimination*. *Best first* ist eine erschöpfende Suche, obwohl die Wahrscheinlichkeit der Untersuchung des ganzen Merkmalssatzes unter Anwendung des Abbruchkriteriums sehr gering ist [Hall, 1998].

Wrapper-Methode

Filter-Methoden arbeiten unabhängig von den Klassifikatoren. Jedoch bestehen die Hauptprobleme darin, dass Filter-Verfahren den Einfluss der Merkmale auf die Leistung des Klassifikators nicht berücksichtigen können. Der beste Merkmalssatz hängt stark von den spezifischen Eigenschaften des Klassifikators ab. Unter Berücksichtigung dieses Faktes verwenden Wrapper-Methoden Klassifikatoren für die Qualitätsbewertung der ausgewählten Merkmale [Aggarwal, 2014].

Eine typische *Wrapper*-Methode besteht aus drei Schritten [Aggarwal, 2014]:

1. die Suche und die Auswahl eines Merkmalsteilsatzes,
2. die Bewertung des ausgewählten Teilsatzes auf Basis der Leistung des Klassifikators und
3. das Wiederholen der ersten zwei Schritte bis die gewünschte Leistung erreicht wird.

Normalerweise wird die Suche in einer von zwei Richtungen durchgeführt: vorwärts vom kleineren zum größeren Merkmalsteilsatz oder rückwärts - vom größeren zum kleineren. In jedem Schritt wird entweder ein Merkmal hinzugefügt oder entfernt. Die Vorwärts-Richtung, bei der man ohne Merkmale anfängt, heißt *forward selection*. Die Rückwärts-Richtung, bei der man mit dem ganzen Merkmalssatz anfängt, heißt *backward elimination* [Witten and Frank, 2011].

In Wrapper-Verfahren funktioniert der vordefinierte Klassifikator als eine Blackbox. Die Komponente für die Suche und Auswahl ergibt einen Merkmalsteilsatz, welcher unter Anwendung des Klassifikators bewertet wird. Das Ergebnis wird zur ersten Komponente für die weiteren Iterationen geschickt. Der beste Merkmalsteilsatz wird zum Anlernen des Klassifikators genutzt. Der resultierende Klassifikator wird danach auf dem unabhängigen Datenteilsatz getestet, welcher nicht während des Trainingsprozesses genutzt wurde [Aggarwal, 2014].

Im Prinzip unterscheiden sich Wrapper-Methoden untereinander durch die eingesetzten Suchmechanismen, weil es möglich ist, jeden Klassifikator anzuwenden. Es existieren diverse Suchstrategien, einige davon werden unten kurz beschrieben.

Bei der ***Sequential forward selection (SFS)*** beginnt die Suche ohne Merkmale. Bei jedem Schritt wird ein Merkmal, welches nicht aus dem aktuellen Merkmalsteilsatz stammt, in den aktuellen Merkmalsteilsatz eingefügt und der resultierende Teilsatz wird bewertet. Am Ende des Schrittes wird das Merkmal mit der besten Bewertung in den Teilsatz eingefügt. Der

Algorithmus wird während einer definierten Schrittanzahl fortgesetzt oder bis keine weitere Verbesserung der Leistung mehr möglich ist [Guyon and Elisseeff, 2006].

Die ***Sequential backward elimination (SBE)*** verwendet am Anfang den ganzen Merkmalssatz. Schrittweise werden Merkmale nacheinander aus dem aktuellen Merkmalsteilsatz gelöscht und die sich ergebenden Merkmalsteilsätze beurteilt. Die Merkmale, welche keinen oder einen negativen Einfluss auf die Leistung haben, werden eliminiert. Der Algorithmus läuft eine voreingestellte Schrittanzahl ab oder bis eine Verschlechterung der Leistung eintritt [Guyon and Elisseeff, 2006, Aggarwal, 2014]. Im Prinzip führt *SFS* schneller als *SBE* zum Ziel, weil *SFS* am Anfang sehr kleine Merkmalsteilsätze berechnet, was oft schneller ist als die Berechnung der fast vollständigen Merkmalssätze mit *SBE*. Anderseits bewertet *SFS* die Merkmale im Zusammenhang mit den Merkmalen, welche schon im Merkmalsteilsatz enthalten sind. Deshalb ist es nicht möglich, die Merkmale, welche nicht selbst, sondern nur in Verbindung mit anderen Merkmalen relevant sind, zu detektieren [Guyon and Elisseeff, 2006].

Die ***Hill-climbing*-Methode** erweitert den aktuellen Merkmalsteilsatz und bewegt sich zum Merkmalsteilsatz mit der höchster Genauigkeit und bricht ab, wenn kein Teilsatz den aktuellen verbessern kann [Aggarwal, 2014].

Die ***Best-first*-Suche** ist eine Methode, welche nicht nur die Suche bei einer Verschlechterung der Leistung abbricht, sondern auch alle vorherigen Qualitätsbewertungen der Merkmale aufbewahrt, welche auf Basis der Leistung eingeordnet worden sind. Daher ist es möglich einen vorherigen Merkmalsteilsatz wieder aufzunehmen [Witten and Frank, 2011].

Manchmal gibt es mehrere vielversprechende Zweige (*branches*) und es ist nicht erwünscht mit der Auswahl eines Weges die weitere Suche zu begrenzen. Es kann wünschenswert sein zu einem anderen als dem zuerst gewählten Zweig zurückzukehren. Die ***Beam*-Suche** bewahrt eine Liste der am interessantesten erscheinenden Zweige (Merkmalsteilsätze) auf, welche noch nicht bewertet wurden. Nach der Untersuchung des einen Zweiges springt der Algorithmus zum anderen Zweig zum Anfang zurück [Guyon and Elisseeff, 2006].

Der **Evolutionäre Algorithmus** (genetischer Algorithmus) ist eine Methode, welche von der biologischen Evolution inspiriert ist: er entwickelt einen guten Merkmalsteilsatz aus der aktuellen Liste von passenden Teilsätzen (Chromosomen) mittels zufälliger Störungen (Mutationen) und Kombinationen (Kreuzungen) [Witten et al., 2016].

Die **simulierte Abkühlung** stellt einen Algorithmus dar, welcher eine Approximation der optimalen Lösung sucht. Der Algorithmus basiert auf der Imitation vom physikalischen Abkühlungsprozess während des Glühens von Metallen. Der Prozess läuft bei einer kontinuierlich absteigenden Temperatur. Die Temperatur ist verantwortlich für die Wahrscheinlichkeit, mit welcher der Algorithmus das Ergebnis verschlechtern kann, um im Weiteren eine bessere Lösung zu finden. Ab der anfänglichen Verteilung sehen die Proben folgendermaßen aus [Witten et al., 2016]:

$$\begin{aligned} {x_1}^{(i+1)} &\sim p(x_1|x_2 = {x_2}^{(i)}, \ldots, x_n = {x_n}^{(i)})^{\frac{1}{t_i}} \\ &\vdots \\ {x_n}^{(i+1)} &\sim p(x_n|x_1 = {x_1}^{(i)}, \ldots, x_{n-1} = {x_{n-1}}^{(i)})^{\frac{1}{t_i}} \end{aligned} \tag{30}$$

wo $p(x_1|x_2)$, $p(x_2|x_1)$ usw. die bedingten Verteilungen sind und die Temperatur mit jeder Iteration abnimmt: $t_{i+1} < t_i$. Jeder Punkt x_n wird nacheinander berechnet und jeder Punkt

ab x_1 kann sich der optimalen Lösung annähern. Bei ausreichender Anzahl der Iterationen konvergiert der Prozess zu einem globalen Minimum [Witten et al., 2016].

Wrapper-Methoden erreichen bessere Ergebnisse als Filter-Methoden, obwohl sie sehr rechenintensiv sind. Ein anderes Problem besteht darin, dass im Prozess der Selektion nur ein Klassifikator genutzt wird, die Ergebnisse sind unvermeidlich vom verwendeten Klassifikator beeinflusst [Aggarwal, 2014].

Embedded-Methoden

Embedded-Methoden unterscheiden sich von anderen Merkmalsselektionsmethoden durch den Lernprozess und die Selektion. Filter-Verfahren enthalten keinen Lernprozess. Wrapper-Methoden nutzen Klassifikatoren für die Qualitätsbewertung der Merkmalsteilsätze ohne Kenntnisse über spezifische Eigenschaften der Funktion des Klassifikators. Die Embedded-Methoden berücksichtigen Modell-Abhängigkeiten und sind weniger rechenintensiv im Vergleich zu Wrapper-Verfahren [Guyon and Elisseeff, 2006].

Entscheidungsbäume werden bei der Trennung der Daten abhängig vom Wert des Merkmals entwickelt. Die passenden Merkmale für die Trennung werden anhand ihrer Bedeutung für die Klassifikation gewählt. Eines der am häufigsten verwendeten Bewertungskriterien ist die Transinformation oder der Informationsgewinn. Häufig reicht der Merkmalsteilsatz, um die Daten komplett zu beschreiben und die Klassifikationsprobleme zu lösen. Weil der Mechanismus für die Merkmalsselektion im Klassifikator eingebunden ist, gehören Entscheidungsbäume zu den Embedded-Verfahren für die Selektion [Guyon and Elisseeff, 2006].

Recursive Feature Elimination (RFE) ist eine Embedded-Methode, welche mittels des Klassifikators nur σ_0 Merkmale anhand von Gewichtskoeffizienten auswählt, um die beste Trennung zwischen den Klassen zu erzielen. Sehr verbreitet ist die Anwendung dieser Methode im Zusammenhang mit Support-Vektor-Maschine (SVM). In dem Fall strebt der Algorithmus einen größten Abstand (*margin*) zwischen den Klassengrenzen an. Dies wird mit einem sogenannten Greedy-Algorithmus gelöst: bei jeder Iteration werden die Merkmale gelöscht, welche am wenigsten den Abstand verringern. Die Berechnung kann beschleunigt werden, indem mehrere Merkmale bei jeder Iteration gelöscht werden. Für die SVM ist der Abstand zwischen Klassen umgekehrt proportional dem Wert w. Zuerst wird der SVM-Klassifikator trainiert und im Ergebnis werden der Vektor α und das Skalar b erhalten. Angesichts des Ergebnisses α wird der Wert w wie folgt berechnet [Guyon and Elisseeff, 2006]:

$$\mathrm{w}^2(\alpha) := \alpha_k \alpha_l y_k y_l K(x_k, x_l) \tag{31}$$

wobei $K(x_k, x_l)$ die Kernfunktion, y_k und y_l die Klassenlabel (kodiert als +1 oder -1) darstellen.

Im Unterschied zu *SBE* trainiert *RFE* nicht den neuen Klassifikator, sondern berechnet die Änderung des Wertes w^2, unter Berücksichtigung der Änderungen des Kernes $k(\cdot)$ nach der Entfernung des Merkmals und in der Annahme, dass der Vektor α feststeht [Guyon and Elisseeff, 2006].

RFE kann auch mit anderen Klassifikatoren arbeiten, welche während des Lernprozesses Gewichtskoeffizienten der Merkmale berechnen können. Vielversprechend ist die Kombination

mit dem Klassifikator ***logistic regression*** [Zhu and Hastie, 2004]. In diesem Fall werden die Gewichtskoeffizienten der Regression als Ranking-Kriterium angewendet.

2.8. Klassifikationsverfahren

Zahlreiche Modelle wurden in der Vergangenheit für die Datenklassifikation entwickelt. Anhand der Datenstruktur unterscheidet man zwei Gruppen an Lernverfahren:

- Unüberwachte maschinelle Lernverfahren und
- Überwachte maschinelle Lernverfahren.

Bei unüberwachten maschinellen Lernverfahren sind die Klassen der Daten nicht vorgegeben und müssen erlernt werden. Dafür gibt es einige Gründe: Sammlung und Markierung der Daten kann für große Datensätze teuer sein; die unmarkierten Daten können zuerst fürs Training angewendet werden und danach werden die Label für die Markierung der Gruppen genutzt; bei mehreren Anwendungen können sich die Eigenschaften der Muster während der Zeit ändern; diese Verfahren können für die Merkmalsselektion angewendet werden; es kann hilfreich sein, die Daten mittels dieser Methode vorübergehend zu untersuchen [Duda et al., 2001].

Die meist verwendeten Algorithmen für das unüberwachte Lernen sind die automatisierte Segmentierung (*Clustering*) und Dimensionsreduktion (Komprimierung der Daten).

Bei überwachten maschinellen Lernverfahren sind die Klassenlabel oder Gewichte für jedes Muster des Trainingsdatensatzes vorgegeben [Duda et al., 2001]. Das Klassifikationsproblem gehört zu dieser Gruppe der Verfahren und besteht darin, dass die Datenstruktur aus Objekten angelernt wird, welche bereits in Gruppen zusammengefasst sind, die den Kategorien oder Klassen entsprechen. Das Lernen dieser Kategorien wird mittels eines Modells erzielt. Das Modell wird zur Bestimmung der Label für unmarkierte und ungesehene Daten angewendet. Die Aufgabe besteht darin, die Klassenlabel für unmarkierte Objekte korrekt vorherzusagen [Aggarwal, 2015].

Die meisten Klassifikationsalgorithmen haben zwei Phasen [Aggarwal, 2015]:

1. die Trainingsphase: bei der das Trainingsmodell aus den Trainingsobjekten aufgebaut wird (zusammenfassendes mathematisches Modell der markierten Objekte des Trainingsdatensatzes) und
2. die Testphase: bei der das aufgebaute Modell zur Bestimmung der Klassenlabel der unmarkierten Test-Objekte angewendet wird.

Es existieren verschiedene Klassifikationsalgorithmen. Einige von diesen sollen hier näher beschrieben werden.

Wahrscheinlichkeitsbasierte Klassifikatoren

Eine verbreitete Subklasse von Klassifikatoren sind die wahrscheinlichkeitsbasierte Verfahren. Diese Verfahren nutzen die statistische Schlussfolgerung, um eine passende Klasse für eine gegebene Instanz zu finden. Zusätzlich zur Bestimmung des Klassenlabels wird auch die

Wahrscheinlichkeit berechnet, mit welcher dieser Datenpunkt zu der einen oder anderen Klasse gehört. Die Klasse mit der höchsten Wahrscheinlichkeit wird als das Label für diesen Datenpunkt gewählt [Aggarwal, 2014].

Naive Bayes

Bayessche Entscheidungstheorie ist grundsätzlicher statistischer Ansatz für Musterklassifikation [Duda et al., 2001].

Der Bayes-Klassifikator basiert auf dem Bayes-Theorem für bedingte Wahrscheinlichkeiten. Mit dem Theorem ist es möglich, die bedingte Wahrscheinlichkeit von einer zufälligen Variablen (das Klassenlabel) auf der Basis von den bekannten Beobachtungen über die Werte von anderen zufälligen Variablen (Merkmale) zu berechnen [Aggarwal, 2015].

D sei ein Datensatz, welcher n Punkte x_i in einem d - dimensionalen Raum hat, und y_i sei Klasse für jeden Punkt, sodass $y_i \in \{c_1, c_2, \ldots, c_k\}$. Ein Bayes-Klassifikator nutzt das Bayes-Theorem, um die Klasse für einen neuen Datenpunkt x vorherzusagen. Das berechnet die a-posteriori-Wahrscheinlichkeit $P(c_i|x)$ für jede Klasse c_i und wählt eine mit der höchsten Wahrscheinlichkeit. Die vorhergesagte Klasse ist wie folgt angegeben [Zaki and Meira Jr., 2014]:

$$\widehat{y} = arg\max_i\{P(c_i|x)\} \tag{32}$$

Das Bayes-Theorem lässt die a-posteriori-Wahrscheinlichkeit durch die Wahrscheinlichkeit und die A-priori-Wahrscheinlichkeit formulieren
[Zaki and Meira Jr., 2014]:

$$P(c_i|x) = \frac{P(x|c_i) \cdot P(c_i)}{P(x)} \tag{33}$$

wobei $P(x|c_i)$ die Wahrscheinlichkeit ist, welche als die Chance von Beobachtung x definiert, in der Annahme, dass die richtige Klasse c_i ist. $P(c_i)$ ist die A-priori-Wahrscheinlichkeit von der Klasse c_i. $P(x)$ ist die Wahrscheinlichkeit von Beobachtung x von jeder aus k Klassen, welche wie folgt berechnet wird [Zaki and Meira Jr., 2014]:

$$P(x) = \sum_{j=1}^{k} P(x|c_j) \cdot P(c_j) \tag{34}$$

Weil $P(x)$ festgestellt für einen Punkt ist, kann die Gleichung 32 umgeschrieben werden [Zaki and Meira Jr., 2014]:

$$\begin{aligned} \widehat{y} &= arg\max_i \left\{ P(c_i|x) \right\} \\ &= arg\max_i \left\{ \frac{P(x|c_i) \cdot P(c_i)}{P(x)} \right\} = arg\max_i \left\{ P(x|c_i) \cdot P(c_i) \right\} \end{aligned} \tag{35}$$

Das bedeutet, dass die vorhergesagte Klasse von der Wahrscheinlichkeit der Klasse in der Annahme der A-priori-Wahrscheinlichkeit abhängt [Zaki and Meira Jr., 2014].

Der Bayes-Klassifikator hat als effektiv für die Text-Klassifikation [McCallum and Nigam, 1998], medizinische Diagnostik [Kelemen et al., 2008] und andere Anwendungen erwiesen [Aggarwal, 2014].

Logistische Regression

Während der Bayes-Klassifikator eine spezielle Form von Wahrscheinlichkeitsverteilung der Merkmale annimmt, entwickelt die Logistische Regression die Wahrscheinlichkeiten der Klassenzugehörigkeit hinsichtlich der Merkmale mit diskriminativer Kraft. Diese Methoden haben verschiedene Modellierungsannahmen, aber beide sind wahrscheinlichkeitsbasierte Verfahren, weil sie eine spezielle Annahme nutzen, um die Merkmale mit einer Wahrscheinlichkeit der Klassenzugehörigkeit zu verbinden. In beiden Fällen werden Parameter des Modells in einer datengesteuerten Weise berechnet. Bei dem einfachsten Modell der logistischen Regression wird der Klassenwert als binär angenommen und wird als $(-1, 1)$ bezeichnet. $\theta = (\theta_0, \theta_1 \dots \theta_d)$ sei ein Vektor von $d + 1$ verschiedener Parameter. θ_i stellt einen Koeffizienten dar, welcher der i-Dimension zugehörig ist, und θ_0 ist ein Offset-Parameter [Aggarwal, 2015]. Die Wahrscheinlichkeit für die Daten $X = (x_1 \dots x_d)$, dass die Klassenvariable Y den Wert $+1$ nimmt, wird wie folgt mit der logistische Regression berechnet [Aggarwal, 2015]:

$$P(Y = 1|X) = f(\theta^T X) = \frac{1}{1 + e^{-\theta^T X}} \tag{36}$$

$$\theta^T X = \theta_0 + \sum_{i=1}^{d} \theta_i x_i \tag{37}$$

wobei $f(x)$ die logistische Funktion ist [Bishop, 2006]:

$$f(x) = \frac{1}{1 + e^{-x}} \tag{38}$$

Die Summe von zwei Wahrscheinlichkeiten ist 1, deshalb die Wahrscheinlichkeit, dass die Klasse Y den Wert -1 nimmt, wie folgt berechnet wird [Aggarwal, 2014, Bishop, 2006]:

$$P(Y = -1|X) = 1 - f(\theta^T X) = \frac{1}{1 + e^{\theta^T X}} \tag{39}$$

Die logistische Regression kann als ein wahrscheinlichkeitsbasierter Klassifikator oder ein linearer Klassifikator gesehen werden. Im Fall der linearen Klassifikatoren wird eine Hyperebene für die Trennung der Klassen angewendet. Die Parameter $\theta = (\theta_0 \dots \theta_d)$ von der logistischen Regression können als Koeffizienten von einer Trennungsebene $\theta_0 + \sum_{i=1}^{d} \theta_i x_i = 0$ dargestellt werden. θ_i ist dann der lineare Koeffizient von der Dimension i und θ_0 ist die Konstante. Der Wert von $\theta_0 + \sum_{i=1}^{d} \theta_i x_i = 0$ wird positiv oder negativ. Das hängt von der Seite der Trennungsebene ab, wo ein Datenpunkt X sich befindet [Aggarwal, 2015].

Im Prinzip wird die Maximum-Likelihood-Funktion $\mathcal{L}(\theta)$ genutzt, um die Parameter der logistischen Regression zu finden. Die Likelihood-Funktion ist ein Produkt der Wahrscheinlichkeiten, dass alle Trainingspunkte die vorhergesagten Labels vom logistischen Modell haben werden [Aggarwal, 2015]. Die gewählten Parameter müssen der folgenden Gleichung entsprechen [Aggarwal, 2014]:

$$\theta \leftarrow \arg\max \mathcal{L}(\theta) \tag{40}$$

wobei $\theta = (\theta_0, \theta_1 \dots \theta_d)$ der Vektor der Parameter ist, welcher berechnet werden muss. Das Modell hat $d + 1$ Anzahl der Parameter für einen d-dimensionalen Raum [Aggarwal, 2014].

Für jeden Datenpunkt gibt es einen Vektor der Parameter $X = x_0, x_1 \dots x_d$ ($x_0 = 1$) und eine beobachtete Klasse $Y = y_k$ [Aggarwal, 2015].

Die Likelihood-Funktion der Parameter kann wie folgt geschrieben werden [Aggarwal, 2014, Aggarwal, 2015, Bishop, 2006]:

$$\begin{aligned}
\mathcal{L}(\theta) &= \prod_{n=1}^{N} P(Y^{(n)} = y_k | X^{(n)}; \theta) \\
&= \prod_{n=1}^{N} \Big(P(Y^{(n)} = 1 | X^{(n)}) \Big)^{Y^{(n)}} \Big(P(Y^{(n)} = 0 | X^{(n)}) \Big)^{1-Y^{(n)}} \\
&= \prod_{n=1}^{N} \Big(\sigma(\theta^T X) \Big)^{Y^{(n)}} \Big(1 - \sigma(\theta^T X) \Big)^{1-Y^{(n)}}
\end{aligned} \tag{41}$$

wobei $Y^{(n)}$ und $X^{(n)}$ entsprechend die beobachtete Werte von Y und X für den n-ten Datenpunkt sind.

Maximierung der Likelihood-Funktion ist ein Äquivalent von Maximierung der logarithmische Likelihood-Funktion [Aggarwal, 2014]:

$$\begin{aligned}
l(\theta) = \log \mathcal{L}(\theta) &= \sum_{n=1}^{N} \log P(Y^{(n)} = y_k | X^{(n)}; \theta) \\
&= \sum_{n=1}^{N} Y^{(n)} (\theta^T X) - log(1 + e^{\theta^T X^{(n)}})
\end{aligned} \tag{42}$$

Es gibt keine geschlossene Lösung zur Maximierung $l(\theta)$ hinsichtlich θ und das Gradientenverfahren wird genutzt.

Gemäß dem Gradientenverfahren und der Ableitungen von jedem θ_i ist es möglich, die Änderung der Gewichte in die Richtung vom Gradient wie folgt zu berechnen [Aggarwal, 2014, Bishop, 2006]:

$$\theta_i = \theta_i + \alpha \sum_{n=1}^{N} \left(Y^{(n)} + \sigma(\theta^T X^{(n)}) \right) X_i^{(n)} \tag{43}$$

wobei α der Lernschritt, eine kleine Konstante, ist. Der Term in Klammern stellt den Vorhersagefehler dar, welcher den Unterschied zwischen der beobachteten $Y^{(n)}$ und vorhergesagten Wahrscheinlichkeit zeigt. Dieser Fehler wird mit dem Wert von $X_i^{(n)}$ multipliziert, um die Größe vom $\theta_i X_i^{(n)}$ - Term zu berücksichtigen [Aggarwal, 2014].

Die logistische Regression wird umfassend in verschiedenen Bereichen genutzt, einschließlich Web, Medizin und Sozialwissenschaften [Aggarwal, 2014]. Das Verfahren kann auch für den Ingenieursbereich angewendet werden, insbesondere für die Vorhersage der Ausfallwahrscheinlichkeit des Systems oder Produktes.

Nearest Neighbor

Die meisten Klassifikationsverfahren bauen ein Modell während der Trainingsphase auf und nutzen dies für Testdaten während der Klassifikationsphase. Das ist auch bekannt als eifriges Lernen (*eager learning*). In Instanz-basierten Lernverfahren gibt es keine reine Trennung zwischen den zwei Phasen [Aggarwal, 2014]. Das Training wird bis zum letzten Schritt verschoben. Solche Klassifikatoren nennt man *"lazy"* [Aggarwal, 2015].

Einer der am meisten verwendeten Klassifikatoren ist der *Nearest-Neighbor-Klassifikator* (NN). In dieser Methode werden die k-nächsten Trainingsinstanzen zum Testpunkt gefunden. Dann wird ein einfaches Modell auf dem Satz er k-nächsten Nachbarn entwickelt, um das Klassenlabel zu definieren. Das dominierende Label zwischen den k-Trainingspunkten wird als das relevante Klassenlabel festgestellt. Das bedeutet, dass das Klassenlabel von den Trainingsobjekten mit den ähnlichsten Merkmalsvektoren dem angegebenen Testobjekt zugeordnet wird [Aggarwal, 2014, Aggarwal, 2015, Runkler, 2012].

Die einfachste Form der Methode ist die Distanzmessung vom Objekt bis zu den k-nächsten Nachbarn. Im Fall der gewichteten Distanz sind alle Nachbarn nicht gleich wichtig. Sei y_i das Klassenlabel für die Instanz i. Die Anzahl der Stimmen $V(j)$ von dem Satz S_k der k-nächsten Nachbarn für das Klassenlabel j wird wie folgt berechnet [Aggarwal, 2014, Friedman et al., 2009]:

$$V(j) = \sum_{i:i \in S_k; y_i = j} f(g_i) \tag{44}$$

wobei g_i ein passendes Unähnlichkeitsmaß ist, z.B. die euklidische oder Mahalabonis-Distanz.

Weil verschiedene Merkmale verschiedene Maßstäbe haben können (z.B. Bewohnerzahl und Alter), ist es notwendig die Merkmale vor der Klassifikation zu normieren.

Dieser Ansatz kann ungeeignet für unausgeglichene Datensätze sein, da dabei einige Klassen weniger Instanzen haben als andere. Zahlreiche Variationen dieses Algorithmus existieren [Aggarwal, 2014]:

- die Auswahl der Distanzfunktion beeinflusst das Verhalten des Klassifikators. Unterschiedliche Distanzfunktionen können besser in verschiedenen Fällen arbeiten.

- Der letzte Schritt der Modellauswahl aus lokalen Testobjekten kann variiert werden, z.B. die Mehrheit der Klassen als das wichtige Ergebnis für die Klassifikation, gewichtete Stimmenmehrheit oder andere Methoden

Der Vorteil des *Nearest-Neighbor*-Klassifikator besteht darin, das der Klassifikator für fast alle Datentypen genutzt werden kann, sobald eine Distanzfunktion die Abstände zwischen Objekten messen kann. Zeitreihen, Textdaten, kategoriale Daten, Multimedia und andere Datentypen können mittels des *NN*-Klassifikators klassifiziert werden [Aggarwal, 2014].

Der wichtigste Nachteil dieses Verfahrens ist die Effizienz der Klassifikation. Die Berechnungszeit hängt linear von der Datensatzgröße ab, was mit immer größeren Datensätze zu einem Problem führt.

Entscheidungsbäume

Die Entscheidungsbäume sind Klassifikationsmethoden, bei welchen der Klassifikationsprozess mittels hierarchischer Entscheidungen an den Merkmalen entwickelt wird, welche in Form einer Baumstruktur angeordnet sind. Die Entscheidungen werden an bestimmten Knotenpunkten des Baums getroffen, welche zum Teilungskriterium (*split criterion*) gehören und abhängig von einem oder mehreren Merkmale aus dem Trainingsdatensatz sind. Das Teilungs-Kriterium teilt den Trainingsdatensatz in zwei oder mehr Teile. Das Ziel ist ein solches Kriterium zu finden, welches die bestmögliche Verringerung der Vermischung der Daten ergibt. Jeder Knotenpunkt im Entscheidungsbaum stellt einen Teilsatz des Datenraums dar, welcher mittels der vorherigen Kombination der Teilungs-Kriterien bestimmt wurde. Der Entscheidungsbaum wird als hierarchische Verteilung der Trainingsobjekte entwickelt [Aggarwal, 2015].

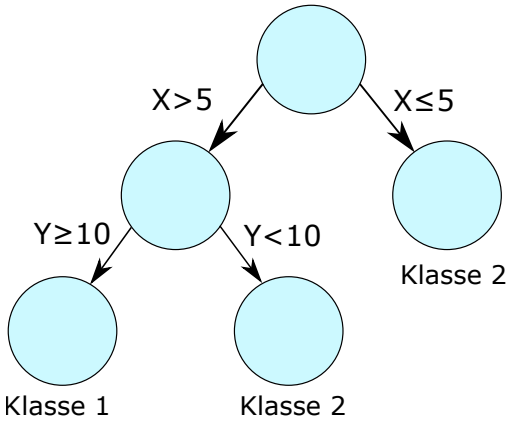

Abbildung 9: Einfacher Entscheidungsbaum

Die Abbildung 9 stellt einen einfachen Entscheidungsbaum dar. Der Wurzelknoten (*root node*) befindet sich obenauf und ist mit anderen Knoten mittels direkter Links verbunden. Diese sind ähnlich miteinander verbunden, solange keine Endknoten (*terminal node*) oder Blattknoten (*leaf node*) erreicht werden, welche selbst keine weitere Verbindung haben. Das ist ein Beispiel von einer allgemeinen Methodologie der Baumentwicklung, welche als **CART** (Classification and Regression Trees) bekannt ist. Laut dieser Methodologie beginnt die Klassifikation des einzelnen Datensatzes im Wurzelknoten, welcher den Wert des Merkmals für den Datensatz prüft. Die verschiedenen Verbindungen entsprechen den verschiedenen

Werten und anhand vom Wert wird eine passende ausgewählt. Das Teilungskriterium in den Knoten muss deutlich die Daten trennen, sodass nur ein Link verfolgt wird. Beim nächsten Schritt wird eine Entscheidung am nachfolgenden Knoten getroffen, welcher als die Wurzel für den Unterbaum betrachtet werden kann. Die Anzahl der Trennungen am Knoten kann beim Entwickler variiert werden und kann unterschiedlich durch den Baum sein. Jeder Entscheidungsbaum kann allein mittels binärer Entscheidungen repräsentiert werden.

So wird der Abstieg über Knoten solange realisiert, bis ein Blattknoten erreicht wird, bei welchem keine weiteren Fragen existieren. Jeder Blattknoten trägt ein Klassenlabel und das Testmodell wird dem Label zugeordnet, dessen Blattknoten erreicht wird [Duda et al., 2001].

Das Hauptprinzip der Baumentwicklung ist es, solche Entscheidungen zu wählen, welche zu einem einfachen und kompakten Baum mit wenigen Knoten führen. Um das zu schaffen, wird ein passendes Kriterium zur Trennung an jedem Knotenpunkt bestimmt, welches die direkt nachfolgenden Daten so einfach wie möglich strukturiert. Die sogenannte Reinheit (*purity*) ist ein Kriterium, welches die Qualität der Trennung misst. Ein Knotenpunkt hat den Reinheitswert von 100%, wenn alle Objekte einer Klasse zugeordnet sind. Es ist besser geeignet, den Gegenwert die Unreinheit (*impurity*) als die Reinheit zu berechnen. Es gibt verschiedene mathematische Maße, welche die Unreinheit beschreiben. Eines der am meisten verwendeten Maße ist das Entropie-Unreinheitsmaß (*entropy impurity*) [Duda et al., 2001]:

$$i(N) = -\sum_{j} P(\omega_j) \log_2 P(\omega_j) \tag{45}$$

wobei $P(\omega_j)$ die Wahrscheinlichkeit von zufälliger Auswahl einer Probe am Knotenpunkt N ist, welche der Kategorie ω_j zugeordnet ist (d.h. Datensatzteil, welcher zur Kategorie ω_j gehört).

Wenn alle Muster in einer Kategorie sind, ist das Entropie-Unreinheitsmaß 0. Der Wert wird positiv und maximal groß in dem Fall, wenn verschiedene Klassen gleich wahrscheinlich sind.

Anhand dieses Wertes wird der Abfall vom Entropie-Unreinheitsmaß am Knotenpunkt N wie folgt berechnet [Duda et al., 2001]:

$$\Delta i(N) = i(N) - P_L i(N_L) - (1 - P_L) i(N_R) \tag{46}$$

wobei N_L und N_R entsprechend der linke und rechte nachfolgende Knoten sind, $i(N_L)$ und $i(N_R)$ ihre Entropie-Unreinheitsmaße; P_L der Datensatzteil am Knotenpunkt N, welcher zum Knotenpunkt N_L beim Entscheidungskriterium T führt.

Für eine Mehrfachtrennung (*multi-way split*) sieht die Gleichung 46 wie folgt aus [Duda et al., 2001]:

$$\Delta i(s) = i(N) - \sum_{k=1}^{B} P_k i(N_k) \tag{47}$$

wobei B der Verzweigungsgrad und P_k eine Fraktion der Daten ist, welche unten zum Knoten N_k geschickt wurde, und $\sum_{k=1}^{B} P_k = 1$.

Diese Optimierung ist lokal. Wie bei verschiedenen Greedy-Algorithmen gibt es die Wahrscheinlichkeit, dass am Ende des Optimierungsprozesses kein globales Optimum erreicht werden kann, d.h. nach dem Training kann nicht der kleinstmögliche Baum entwickelt werden.

Ein anderes Problem bei der Entwicklung eines Baumes ist die Überanpassung des Entwicklungsprozesses (*overfitting*), welche beim unbegrenzten Wachsen entstehen kann. Es existieren verschiedene Methoden, um dieses Problem zu lösen. Ein traditioneller Ansatz besteht in der Anwendung des Kreuzvalidierungsverfahrens (*cross-validation*). Eine andere Methode ist die Schwellwertmethode (*threshold*). Diese hat ihre Vorteile in der Anwendung des ganzen Datensatzes fürs Training im Gegensatz zur Kreuzvalidierung. Der Nachteil besteht darin, dass es schwer ist, den Zusammenhang zwischen einem Schwellwert und der Klassifikationsleistung zu bestimmen. Eine einfache Lösung ist aufzuhören, wenn der Knoten eine bestimmte Anzahl an Datenpunkten oder Prozent vom Datensatz repräsentiert.

Eine Alternative für diese Methoden ist das Festlegen eines Abbruchkriteriums anhand der statistischen Signifikanz der Verringerung der Unreinheit. Für alle Trennungskriterien wird der Chi-Quadrat-Test durchgeführt. Wenn das Kriterium die Unreinheit nicht wesentlich verringert, wird die Trennung abgebrochen.

Die Anwendung des Abbruchkriteriums für die Trennung hat einen Nachteil, ihr fehlt die Möglichkeit des Vorausschauens (s.g. Horizonteffekt (*horizon effect*)). Die Bestimmung des Knotenpunktes N wird nicht von der Entscheidung am nachfolgenden Knoten beeinflusst. Im Fall des Abbruchkriteriums kann der Knotenpunkt N als Blatt bestimmt werden, dadurch wird die Möglichkeit für eine Trennung im nachfolgenden Knoten ausgeschlossen, so dass der Prozess zu früh für das Erreichen einer bestmöglichen Erkennungsgenauigkeit abgebrochen werden kann. Das Abbruchkriterium beeinflusst den Lernalgorithmus so, dass die Bäume mit hoher Unreinheit an den Wurzelknoten bevorzugt werden.

Die Alternative für die Anwendung eines Abbruchkriteriums ist das sogenannte *pruning*. Es lässt den Baum völlig wachsen, bis die Blattknoten eine minimale Unreinheit aufweisen. Alle Blattknoten danach werden für die Beschneidung berücksichtigt. Bei einer kleinen Unreinheit werden diese Knoten abgeschnitten und vorhergehende Knoten als Blattknoten angegeben. Diese Knoten können im Weiteren auch eliminiert werden. Das ist im Prinzip eine umgekehrte Trennung. Nach dem Pruning werden Blattknoten oft auf verschiedenen Niveaus liegen, so dass der Baum unausgeglichen wirkt.

Der Vorteil des Prunings ist die Eliminierung des Horizonteffektes. Auch bei dieser Methode werden keine Datenteile für die Kreuzvalidierung vorgehalten, es nutzt hingegen alle Information im Trainingssatz. Offensichtlich ist diese Methode rechenintensiver als die abgebrochene Trennung und führt im Fall eines großen Datensatzes zu unzulässig hohen Rechenzeiten. Für kleine Datensätze ist die Rechenintensität dieser Methode gering und hier wird das Pruning der abgebrochenen Trennung vorgezogen [Duda et al., 2001].

Grundsätzlich können alle Entscheidungsbaum-Algorithmen mit oben beschriebenen Komponenten und Verfahren enthalten sein. In der Realität nutzen die Algorithmen verschiedene Abbruchkriterien, Unreinheits-Werte, usw. [Duda et al., 2001].

Es existieren zahlreiche Realisierungen an Entscheidungsbäumen, die am meisten verwendet werden *C4.5*, *Iterative Dichotomiser 3 (ID3)* und der oben beschriebene *CART*-Algorithmus.

ID3 ist ein einfacher Entscheidungsbaum-Algorithmus. Diese Methode involviert reellwertige Variablen, welche zuerst auf mehrere Intervalle aufgeteilt werden. Jedes Intervall wird

als ungeordnetes Attribut behandelt. Jede Trennung hat einen eigenen Verzweigungsfaktor B_j, wobei B_j die Anzahl der für die Trennung ausgewählten Attribute der Variablen j ist. Für die Berechnung wird der skalierte Wert, der *gain ratio impurity* heißt, aus der Gleichung 47 verwendet [Duda et al., 2001]:

$$\Delta i_B(s) = \frac{i(s)}{-\sum_{k=1}^{B} P_k log_2(P_k)} \tag{48}$$

Die Anzahl von Ebenen bei solchen Entscheidungsbäumen sind gleich der Anzahl der Eingangsvariablen. Der Algorithmus arbeitet iterativ, bis alle Knoten rein sind oder es gibt keine weiteren Variablen. Obwohl das originale Konzept von *ID3* kein Pruning enthält, ist es möglich, diese Methode zu implementieren [Duda et al., 2001].

C4.5 ist die meist verwendete Klassifikationsmethode bei Entscheidungsbäumen. In *C4.5* werden reelwertige Variablen gleich wie in *CART* behandelt. Dieser Algorithmus nutzt ein Pruning anhand der statistischen Signifikanz der Trennung. Der Unterschied zwischen *C4.5* und ***CART*** liegt in der Arbeit mit fehlenden Attributswerten. Während des Trainingsprozesses gibt es keinen speziellen Platz für diese Werte und es werden weder Vorberechnungen noch Ersatztrennungen gemacht, fehlende Werten werden nicht bei der Berechnung von *impurity* an den Knoten verwendet. Ein anderer Unterschied liegt in der Berechnung von *impurity*: *C4.5* nutzt wie *ID3* den Informationsgewinn (Gleichung 47) und *CART - Gini index* [Duda et al., 2001]:

$$i(N) = 1 - \sum_j P^2(\omega_j) \tag{49}$$

Entscheidungsbäume können als einfache Klassifikatoren zur Definition der effektiven Lernalgorithmen verwendet werden. Sie sind normalerweise schnell zu berechnen und einfach zu interpretieren [Mohri et al., 2012].

Random Forest

Ein Ansatz für die Verringerung der Datenvarianz besteht darin, dass viele Berechnungen des Wertes gemittelt werden. Im Fall des Entscheidungsbaums können mehrere Bäume auf den verschiedenen Teilsätzen, welche zufällig ausgewählt werden, trainiert werden und dann wird ein Ensemble an Bäumen berechnet [Murphy, 2012]:

$$f(x) = \sum_{i=1}^{N} \frac{1}{N} f_i(x) \tag{50}$$

wobei f_i der i-te Baum und N Anzahl der Bäume im Ensemble ist.

Diese Methode heißt *Bagging* [Breiman, 1996]. Eine einfache Wiederdurchführung des gleichen Algorithmus auf verschiedenen Teildatensätzen kann hoch korrelierte Aussagen

produzieren, was die mögliche Verringerung der Varianz begrenzt. Dieser Ansatz ist bekannt als Random Forest (RF) und versucht verschiedene Bäume zu dekorrelieren, dafür werden ein Merkmalsteilsatz und ein Datenteilsatz zufällig ausgewählt [Murphy, 2012].

Die Entscheidung über die Klassenzuordnung wird in dem Fall anhand des ermittelten Ergebnisses aller Bäume getroffen. Für ein gegebenes Ensemble an Klassifikatoren $f_1(x), f_2(x) \dots f_N(x)$ und einen zufälligen Vektor Y, X aus dem Trainingssatz, berechnet sich die Randfunktion wie folgt [Breiman, 2001]:

$$margin(X,Y) = \frac{1}{N}\sum_{i=1}^{N} I(f_i(X) = Y) - \max_{j \neq Y}\{\frac{1}{N}\sum_{i=1}^{N} I(f_i(X) = j)\} \tag{51}$$

wobei $I(\cdot)$ die Indikatorfunktion ist.

Die Funktion ist ein Maß, welches den Abstand zwischen der mittleren Stimmenzahl von X, Y für die richtige Klasse und der mittleren Stimmenzahl für alle andere j Klassen anzeigt. Je größer der Abstand ist, desto mehr Vertrauen in die Klassifikation liegt vor.

Random Forest ist der effektivste Entscheidungsbaumklassifikator. Wegen des Gesetzes der großen Zahlen leidet RF nicht unter der Überanpassung (*overfitting*) [Breiman, 2001].

Support-Vektor-Maschine

Die Support-Vektor-Maschine (englisch - *support vector machine*)(SVM) wurde zu Beginn für binäre Klassifikationen numerischer Daten entwickelt. Das binäre Problem kann auch als Mehrklassenproblem mittels einiger Methoden erweitert werden. Mit kategorischen Werten können diese Klassifikatoren auch durch die Verwandlung in binäre Daten arbeiten.

Wie bei linearen Modellen trennt die SVM mittels Hyperebenen (siehe Gleichung 9), welche als Grenze zwischen zwei Klassen gilt.

Hard-margin SVM

Im Fall linear trennbarer Klassen existiert eine unendliche Anzahl an Hyperebenen, welche die Klassen trennen können [Aggarwal, 2015]. Es ist möglich, die parallelen Hyperebenen zu finden, welche die Trainingsdaten von beiden Seiten berühren und keine anderen Datenpunkte dazwischen haben. Die Trainingsdatenpunkte sind Stützvektoren (*support vectors*) und der Abstand zwischen Hyperebene und Trennungsebene ist der Randabstand (*margin*). Die Trennungsebene (*decision boundary*) befindet sich genau in der Mitte zwischen den beiden Hyperebenen, um eine bestmögliche Klassifikation zu erreichen [Bishop, 2006]. Die Hyperebene, bei welcher der Mindestabstand zu den Trainingsdatenpunkten möglichst maximal ist, ist die robusteste für eine korrekte Klassifikation. Im Fall der SVM sind Optimierungsprobleme zu lösen und solche Hyperebenen mit dem Gedanken an den Randabstand zu finden [Aggarwal, 2015].

Die optimalen Koeffizienten können durch die Lösung der Optimierungsaufgabe erhalten werden.

Die $(d + 1)$ Koeffizienten für w und b müssen optimiert werden, um den maximalen Randabstand für die Trennung zweier Klassen zu erreichen. Alle Datenpunkte x_i (wobei $i = 1, \dots, n$) mit $y_i = +1$ liegen auf einer Seite der Hyperebene und genügen der Bedingung $\mathrm{w}^T \cdot \mathrm{x}_i + b \geq 0$, alle Punkte mit $y_i = -1$ liegen auf der anderen Seite und genügen $\mathrm{w}^T \cdot \mathrm{x}_i + b \leq 0$ [Aggarwal, 2015]:

$$\begin{aligned} \mathrm{w}^T \cdot \mathrm{x}_i + b \geq 0 \qquad & \forall i : y_i = +1 \\ \mathrm{w}^T \cdot \mathrm{x}_i + b \leq 0 \qquad & \forall i : y_i = -1 \end{aligned} \tag{52}$$

Es kann angenommen werden, dass die Trennungsebene $\mathrm{w}^T \cdot \mathrm{x} + b = 0$ in der Mitte zwischen zwei Hyperebenen liegt. Deshalb können die Hyperebenen mit dem Parameter v, welcher die Distanz dazwischen einstellt, ausgedrückt werden [Aggarwal, 2015]:

$$\begin{aligned} \mathrm{w}^T \cdot \mathrm{x} + b = +v \\ \mathrm{w}^T \cdot \mathrm{x} + b = -v \end{aligned} \tag{53}$$

Bei der optimalen Skalierung der Variablen w und b ist es möglich den Wert v als 1 anzusetzen. Dann werden zwei Trennungshyperebenen wie folgt bestimmt [Aggarwal, 2015]:

$$\begin{aligned} \mathrm{w}^T \cdot \mathrm{x} + b = +1 \\ \mathrm{w}^T \cdot \mathrm{x} + b = -1 \end{aligned} \tag{54}$$

Diese Bedingungen werden als Randbedingungen bezeichnet. Zwei Hyperebenen trennen den Datenraum in drei Bereiche. Es wird angenommen, dass keine Datenpunkte dazwischen liegen und alle Trainingsdatenpunkten für jede Klasse auf einer der zwei anderen Regionen abgebildet werden (die sogenannte *Hard-margin SVM*) (Abbildung 10). Das kann als punktweise geltende Bedingungen auf den Trainingsdatenpunkten definiert werden [Aggarwal, 2015]:

$$\begin{aligned} \mathrm{w}^T \cdot \mathrm{x}_i + b \geq +1 \qquad & \forall i : y_i = +1 \\ \mathrm{w}^T \cdot \mathrm{x}_i + b \leq -1 \qquad & \forall i : y_i = -1 \end{aligned} \tag{55}$$

Der Abstand zwischen den Hyperebenen ist der normalisierte Unterschied zwischen den Konstanten, wo der Normierungsfaktor euklidischer Abstand (L^2-Norm) wie folgt definiert ist [Aggarwal, 2015]:

$$\|\mathrm{w}\| = \sqrt{\sum_{i=1}^{d} \mathrm{w}_i^2} \tag{56}$$

Weil der Unterschied zwischen Konstanten von Hyperebenen 2 ist, ist die Distanz $\frac{2}{\|\mathrm{w}\|}$ und der Randabstand entsprechend $\frac{1}{\|\mathrm{w}\|}$. Dieser Abstand muss maximiert werden bzw., was dasselbe ist, $\|\mathrm{w}\|$ bzw. $\frac{\|\mathrm{w}\|^2}{2}$ müssen minimiert werden [Aggarwal, 2015] [Zaki and Meira Jr., 2014]:

$$\begin{aligned} \textbf{Zielfunktion:} \quad & \min_{\mathrm{w},b} \{\frac{\|\mathrm{w}\|^2}{2}\} \\ \textbf{Lineare Bedingungen:} \quad & y_i(\mathrm{w}^T \cdot \mathrm{x}_i + b) \geq 1 \qquad \forall x_i \in \mathbf{D} \end{aligned} \tag{57}$$

wo **D** ein Datensatz ist.

Hierbei handelt es sich um konvexe quadratische Optimierungsprobleme, weil die quadratische Funktion $\frac{\|\mathrm{w}\|^2}{2}$ vorbehaltlich eines Satzes der linearen Gleichungen 52 bzw. 55 minimiert werden muss [Aggarwal, 2014].

Das Problem der nicht linearen Programmierung kann mittels der Methode des Lagrange-Relaxationansatzes (*Lagrangian relaxation*) gelöst werden. Die Idee besteht darin, dass ein n-dimensionaler nicht negativer Satz an Lagrange-Multiplikatoren $\alpha = (\alpha_1 ... \alpha_n) \geq 0$ mit verschiedenen Bedingungen aus der Gleichung 55 assoziiert wird. Der Multiplikator α_i entspricht der Randbedingung des i-ten Trainingsdatenpunktes. Die Bedingungen sind dann erfüllt und die Zielfunktion wird mit der Einführung der Lagrange-Strafe für die Bedingungsunterbrechung vergrößert [Aggarwal, 2015]:

$$L_p = \frac{\|\mathrm{w}\|^2}{2} - \sum_{i=1}^{n} \alpha_i \{y_i(\mathrm{w}^T \mathrm{x}_i + b) - 1\} \tag{58}$$

L_p soll hinsichtlich w und b minimiert und hinsichtlich α_i maximiert werden.

Wenn die Ableitung von L nach w und b auf Null gesetzt wird, erhält man [Zaki and Meira Jr., 2014]:

$$\frac{\partial}{\partial \mathrm{w}} L = \mathrm{w} - \sum_{i=1}^{n} \alpha_i y_i x_i = 0 \qquad \Longrightarrow \qquad \mathrm{w} = \sum_{i=1}^{n} \alpha_i y_i x_i \tag{59}$$

$$\frac{\partial}{\partial b} L = -\sum_{i=1}^{n} \alpha_i y_i = 0 \qquad \Longrightarrow \qquad \sum_{i=1}^{n} \alpha_i y_i = 0 \tag{60}$$

Die Gleichungen 59, 60 geben eine Ahnung vom optimalen Gewichtsvektor w. Die Gleichung 59 bedeutet, dass w als eine lineare Kombination von Datenpunkten x_i und Lagrange-Multiplikatoren $\alpha_i y_i$ (dienen als Koeffizienten) ausgedrückt werden kann. Die Gleichung 60 bedeutet, dass die Summe aller Lagrange-Multiplikatoren $\alpha_i y_i$ Null sein muss. Beim Verwenden dieser Gleichungen in der Gleichung 58 erhält man die doppelte Lagrange-Zielfunktion, welche nur in Form von Langrange-Multiplikatoren angegeben wird [Zaki and Meira Jr., 2014]:

$$\begin{aligned} L_{dual} &= \frac{1}{2}\mathrm{w}^T\mathrm{w} - \mathrm{w}^T \underbrace{\left(\sum_{i=1}^{n} \alpha_i y_i x_i \right)}_{\mathrm{w}} - b \underbrace{\sum_{i=1}^{n} \alpha_i y_i}_{0} + \sum_{i=1}^{n} \alpha_i \\ &= -\frac{1}{2}\mathrm{w}^T\mathrm{w} + \sum_{i=1}^{n} \alpha_i \\ &= \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j x_i^T x_j \end{aligned} \tag{61}$$

Somit wird die doppelte Zielsetzung wie folgt angegeben [Zaki and Meira Jr., 2014]:

$$\textbf{Zielfunktion:} \quad \max_{\alpha} \; L_{dual} = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j x_i^T x_j \tag{62}$$

$$\textbf{Lineare Bedingungen:} \quad \alpha_i \geq 0, \forall i \in \mathbf{D}, \quad \text{and} \quad \sum_{i=1}^{n} \alpha_i y_i = 0 \tag{63}$$

L_{dual} ist ein konvexes Problem der quadratischen Programmierung. Dieses Problem kann unter Verwendung standardmäßiger Optimierungsmethoden gelöst werden [Zaki and Meira Jr., 2014].

Für die Klassifikation der Instanz u wird das Klassenlabel $\widehat{y}(u)$ hinsichtlich der Lagrange-Multiplikatoren (Gleichung 59) definiert als [Aggarwal, 2015]:

$$\widehat{y}(z) = sign\{\mathrm{w}^T \cdot u + b\} = sign\left\{ \left(\sum_{i=1}^{n} \alpha_i y_i x_i u \right) + b \right\} \tag{64}$$

wobei $sign(\cdot)$ eine Funktion ist, welche $+1$ zurückgibt, wenn ihr Argument positiv ist, und -1, wenn es negativ ist [Zaki and Meira Jr., 2014].

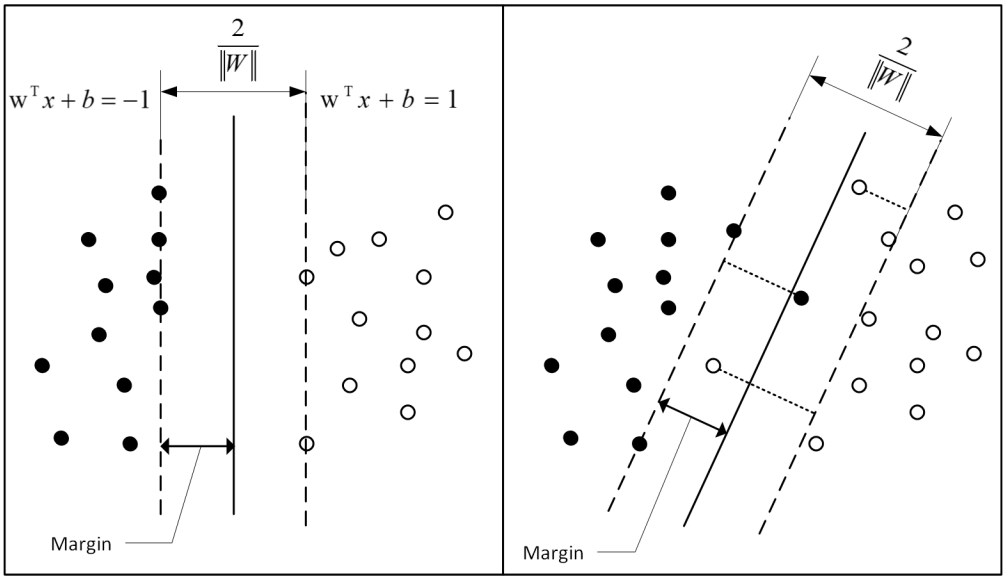

Abbildung 10: SVM mit Hard- und Soft-Margin (in Anlehnung an [Zaki and Meira Jr., 2014])

Soft-margin SVM

Ideal trennbare Daten kommen in der Regel nur für künstliche Daten vor, reale Daten weisen zumeist nichtlineare Eigenschaften auf, obwohl Datensätze annähernd trennbar sein können. Die zwei Hyperebenen trennen die Mehrheit an Datenpunkten aber nicht alle (die sogenannte *Soft-Margin SVM*, Abbildung 10) [Aggarwal, 2015]. Um die nicht trennbaren Punkten zu bestrafen, wird die Schlupfvariable ξ verwendet. $\xi_i \geq 0$, wobei $i = 1, \ldots, n$ eine

Schlupfvariable für jeden Datenpunkt ist [Cortes and Vapnik, 1995]. Die Bedingungen für die Hyperebenen können dann wie folgt beschrieben werden [Aggarwal, 2015]:

$$\begin{aligned} \mathrm{w}^T \cdot \mathrm{x}_i + b \geq +1 - \xi_i \qquad & \forall i : y_i = +1 \\ \mathrm{w}^T \cdot \mathrm{x}_i + b \leq -1 + \xi_i \qquad & \forall i : y_i = -1 \end{aligned} \tag{65}$$

ξ kann als Abstand vom Datenpunkt zu den Trennungsebenen interpretiert werden. Es ist nicht erwünscht mehrere Datenpunkte mit einem positiven Wert zu haben, deshalb wird ein solcher Verstoß durch $C \cdot \xi_i^r$ bestraft. C und r sind einstellbare Parameter der Weichheit des Modells. Ein kleiner Wert C entspricht einem breiten Randabstand (*relaxed margin*), und ein größerer wird den Trainingsdatenfehler minimieren und entspricht einem engen Randabstand (*narrow margin*). Wird C ausreichend groß eingestellt, werden keine Datenfehler zugelassen, was dieselbe Einstellung als Hard-margin SVM hat. Eine gängige Wahl für den Parameter r ist der Wert 1, was einer sogenannten Hinge-Verlustfunktion (*hinge loss*) entspricht. Die Zielfunktion für die $Soft - margin$ SVM wird wie folgt definiert [Zaki and Meira Jr., 2014, Aggarwal, 2015, Bishop, 2006]:

$$\textbf{Zielfunktion:} \quad \min_{\mathrm{w},b,\xi_i} \left\{ \frac{1}{2} \mathrm{w}^T \mathrm{w} + C \sum_{i=1}^{n} \xi_i^r \right\} \tag{66}$$

$$\textbf{Lineare Bedingungen:} \quad \xi_i \geq 0, \quad y_i(\mathrm{w}^T x_i + b) \geq 1 - \xi_i \quad \forall x_i \in \mathbf{D} \tag{67}$$

Diese Formeln bestehen aus zwei Teilen: der Norm des maximalen Abstandes zur Hyperebene und der Verlustgröße wegen einer Verletzung des Modells. Den Formelteil des Maximums des Randabstandes bezeichnet man auch als Regularisierungsfaktor (*regularization term*), welcher die Komplexität des Modells hinsichtlich der Verletzung des Modells einstellt. Der Regularisierungsfaktor hat auch eine andere Funktion, er balanciert die Genauigkeit des Modells und den Randabstand aus [Aggarwal, 2014].

Im Gegensatz zur Maximierung des Abstandes im Fall der trennbaren Daten ist es notwendig, auch die Verluste $C \sum_{i=1}^{n} \xi_i^r$ zu minimieren unter Wahrung eines großen Abstandes. Das sind zwei Ziele und es gibt mehrere Wege das auszugleichen. Zum Beispiel ist es möglich, einen kleineren Randabstand nur mit einer verletzenden Instanz zu haben. Alternativ kann ein größerer Randabstand mit mehreren verletzenden Datenpunkte vorkommen.

Kernel SVM

Das lineare Modell kann für einige Daten ungeeignet sein. So sind in Abbildung 11 (links) beispielhaft nicht linear trennbare Daten dargestellt. In diesem Fall kann ein linearer Klassifikator maximal eine Erkennungsrate von 50% erreichen, was nicht besser als eine zufällige Vorhersage ist. Wenn allerdings eine Merkmalstransformation $\phi(x)$ (Gleichung 68 [Aggarwal, 2014]) angewendet wird, liegen die Daten in einem mehrdimensionalen Merkmalsraum plötzlich linear trennbar vor (Abbildung 11 (rechts)).

$$\phi(x) = [x_1 x_1 \quad x_2 x_2]^T \tag{68}$$

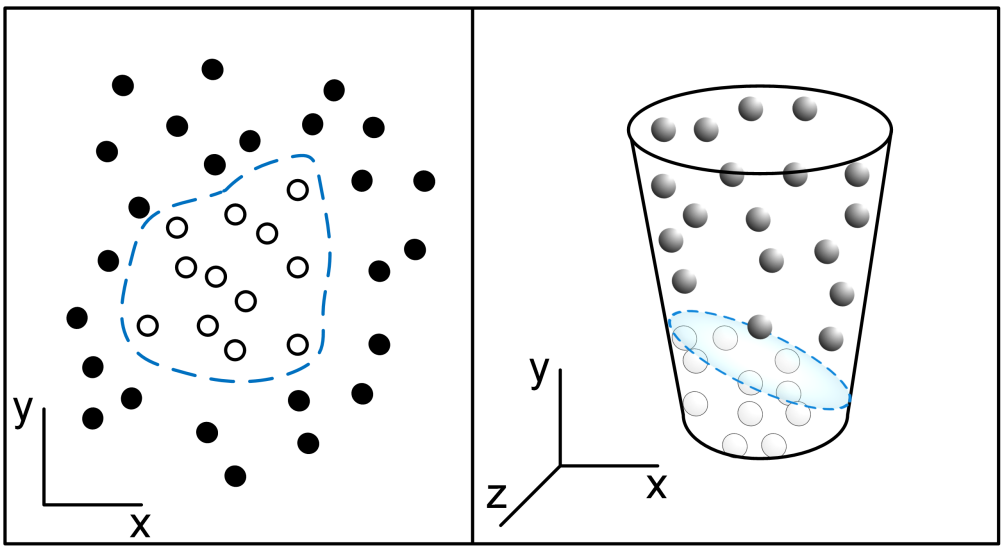

Abbildung 11: Nicht linear trennbare Daten im originalen Merkmalsraum (links) und dieselben Daten linear trennbar durch Kernel-Trick im mehrdimensionalen Merkmalsraum (rechts)

Die Idee besteht darin, dass die originalen d-dimensionalen Daten x_i in Daten $\phi(x_i)$ des mehrdimensionalen Merkmalsraums durch eine nicht lineare Transformation ϕ umgewandelt werden. Angesichts der größeren Flexibilität ist es wahrscheinlich, dass die Daten $\phi(x_i)$ im neuen Merkmalsraum linear trennbar sind. Diese lineare Trennebene im mehrdimensionalen Raum entspricht der nicht linearen Ebene im originalen Merkmalsraum [Zaki and Meira Jr., 2014].

Das Klassenlabel für den neuen Punkt u kann wie folgt berechnet werden [Zaki and Meira Jr., 2014]:

$$\begin{aligned} \widehat{y} &= sign(\mathrm{w}^T \phi(u) + b) \\ &= sign\left(\sum_{\alpha_i > 0} \alpha_i y_i \phi(x_i)^T \phi(u) + b \right) \\ &= sign\left(\sum_{\alpha_i > 0} \alpha_i y_i K(x_i, u) + b \right) \end{aligned} \tag{69}$$

Aus der Gleichung 69 folgt, dass nur das Skalarprodukt im originalen Merkmalsraum für die Berechnung $\widehat{y}$ notwendig ist. Alle Operationen können durch die Kernfunktion K (Gleichung 70 [Zaki and Meira Jr., 2014]) durchgeführt werden.

$$K(x_i, x_j) = \phi(x_i)^T \phi(x_j) \tag{70}$$

Diese Methode wird als sogenannter Kernel-Trick (*kernel trick*) bezeichnet. Dadurch kann

eine nicht lineare Kernfunktion für eine nicht lineare Klassifikation im Eingaberaum genutzt werden [Zaki and Meira Jr., 2014].

Es existieren mehrere Kernfunktionen. Die am häufigsten Gebrauchten werden nachfolgend dargestellt.

Polynomkern *(polynomial kernel)*. Für jede Konstante $a > 0$, ist ein Polynomkern K des Grades $d \in N$ über R^N (ein Merkmalsraum mit N-Dimensionen) wie folgt definiert [Mohri et al., 2012]:

$$\forall x, x' \in R^N, \quad K(x, x') = (x \cdot x' + a)^d. \tag{71}$$

Der Polynomkern bildet den Eingaberaum in einem höherdimensionalen Raum der Größe $\binom{N+d}{d}$ ab [Mohri et al., 2012].

Gauß-Kern *(gaussian kernel)*. Für jede Konstante $\sigma_{RBF} > 0$, ist der Gauß-Kern oder die radiale Basisfunktion (RBF) definiert als [Mohri et al., 2012]:

$$\forall x, x' \in R^N, \quad K(x, x') = exp\bigg(- \frac{||x' - x||^2}{2\sigma_{RBF}^2} \bigg) \tag{72}$$

Sigmoid-Kern *(sigmoid kernel)*. Für jede Konstante $a_1, a_2 \geq 0$, ist der Sigmoid-Kern [Mohri et al., 2012]:

$$\forall x, x' \in R^N, \quad K(x, x') = tanh(a_1(x \cdot x') + a_2) \tag{73}$$

Die Anwendung des Sigmoid-Kerns führt zu einem Algorithmus, welcher eng mit Algorithmen auf der Basis einfacher neuronaler Netze verbunden ist, welche ebenfalls oft durch eine Sigmoidfunktion definiert sind [Mohri et al., 2012].

Neuronale Netze

Das künstliche neuronale Netz (KNN, *artificial neural network (ANN)*) ist ein Algorithmus, welcher aus individuellen, miteinander verbundenen Neuronen (in verschiedenen Quellen findet man sie auch unter den Namen: Einheiten, Knoten, Units bezeichnet) aufgebaut und dem menschlichen Nervensystem ähnlich ist [Heaton, 2015].

Das künstliche Neuron (Abbildung 12 (rechts)) ist eine Berechnungseinheit, welche den Satz der Eingänge mittels zweier Funktionen ξ und h in einen Ausgang verwandelt. Die Neuronen sind miteinander bidirektional oder unidirektional verbunden [Kruse et al., 2015]. Die Stärke der Verbindung zwischen Neuronen wird durch das Gewicht w definiert. Während des Lernprozesses werden diese Gewichte angepasst [Aggarwal, 2014, Aggarwal, 2015, Bishop, 2006].

Das Neuron nimmt die Eingangsdaten und vereinigt die mittels einer Netzeingabefunktion $\xi(\cdot)$ in einem Wert, welcher Aktivierung (*net activation*) heißt [Aggarwal, 2014]:

$$net = \xi(x, w) \tag{74}$$

Normalerweise stellt eine Netzeingabefunktion eine Summe, Distanz oder einen Kernel dar.

Dann wird die Aktivierung mittels einer Aktivierungsfunktion (*activation function*) $h(\cdot)$ in die Ausgabe des Neurons z verwandelt [Bishop, 2006]:

$$z = h(net) \tag{75}$$

Die Auswahl der Aktivierungsfunktion hängt von der Art der Daten ab. Verschiedene Aktivierungsfunktionen können auf die Aktivierungen von Neuronen in allen Schichten des Netzes angewendet werden [Duda et al., 2001]. Radiale Basisfunktionen, Signumfunktionen, Schwellwertfunktionen, lineare Funktionen, Sigmoidfunktion, Polynomfunktionen u.a kommen in Frage [Aggarwal, 2014]. Das hängt auch von dem Platz des Neurons im Netz ab.

Das Netz ordnet Neuronen in hierarchischen Schichten, wobei alle Schichten eine spezielle Funktionalität bieten, wie z.B. Eingangsdaten empfangen, Zwischenprodukt berechnen oder ein Ergebnis ausgeben. Neuronen in der Eingangsschicht führen keine Berechnungen durch und nehmen einfach Eingangsdaten ins Netz auf, deshalb wird diese Schicht in der Schichtenanzahl nicht miteinbezogen, sodass ein einlagiges Netz aus einer Eingangsschicht und einer Ausgangsschicht besteht. Es existieren verschiedene Architekturen neuronaler Netze: von einfachen einlagigen bis mehrlagigen komplexen Netzen [Aggarwal, 2014, Aggarwal, 2015, Bishop, 2006].

Einlagiges Perzeptron

Das einlagige Perzeptron (*single layer perceptron (SLP)*) ist ein Beispiel von einlagigen Netzen, welches aus zwei Schichten besteht: einer Eingangsschicht und einer Ausgangsschicht. Es ist eine der ersten und einfachsten Architekturen des neuronalen Netzes. Die Ausgangsschicht besteht aus einem linearen Schwellwertelement (*linear threshold unit*), welches die Eingangswerte x_i von Neuronen aus der Eingangsschicht bekommt. Die Stärke der Verbindung wird mittels der Gewichte w_i definiert. Das Netz nimmt die Eingangsdaten in die Eingangsschicht und trägt das direkt zur Ausgangsschicht. Als die Netzeingabefunktion nutzt dieses lineare Schwellwertelement die gewichtete Summe von Eingänge [Aggarwal, 2014, Bishop, 2006]:

$$net = \xi(x, w) = \sum_{i=1}^{D} w_i x_i + w_0 \tag{76}$$

wo D - Anzahl der Neuronen in der Eingangsschicht des Netzes ist, w_0 ist Bias, welches die Trennungsebene (Hyperebene) verschiebt und unabhängig von Eingängen ist.

Das Bias kann in den Satz der Gewichte aufgenommen werden, wenn eine Variabel x_0 mit dem Wert $x_0 = 1$ eingeführt wird. Die Gleichung 76 wird dann wie folgt geschrieben [Bishop, 2006]:

$$net = \xi(x, w) = \sum_{i=0}^{D} w_i x_i \tag{77}$$

Die Aktivierungsfunktion des linearen Schwellwertelementes ist eine Signum- oder Treppenfunktion, sodass die Aktivierung in die Ausgabe des Netzes z wie folgt verwandelt wird [Bishop, 2006]:

$$z = h(net) = sign(net) = sign\left(\sum_{i=0}^{D} w_i x_i\right) = \begin{cases} 1 & \text{if } net \geq 0 \\ -1 & \text{if } net < 0 \end{cases} \tag{78}$$

Der Wert z_i stellt die Prognose des Klassenlabels für die Variable x_i dar. Es ist wünschenswert, solche Gewichte zu wählen, dass die Prognose z_i gleich dem originalen Klassenlabel z' für so viele Trainingspunkten x_i wie möglich funktioniert. Der Fehler der Prognose $(z_i - z_i')$ kann jeden der Werte $-2, 0, 2$ haben. 0 entspricht der Situation, wenn die Prognose richtig ist. Das Ziel des Trainings ist, die Gewichte und Biases so einzustellen, dass die Prognose so nah wie möglich an den originalen Klassenvariablen ist. Der Algorithmus nimmt die Eingangsdaten x_i nacheinander ins Netz, um die Prognose z_i zu erstellen. Die Gewichte werden dann auf der Basis des Fehlers geändert, sodass der Gewichtsvektor w nach t Iterationen wie folgt aktualisiert wird: [Aggarwal, 2015]

$$w^{t+1} = w^t + \eta(z_i' - z_i)x_i \tag{79}$$

Der Parameter η steuert die Lernrate des neuronalen Netzes. Der Algorithmus wiederholt die Optimierung aller Trainingspunkte und passt die Parameter des Netzes schrittweise an, bis die Konvergenz erreicht wird. Jeder Punkt kann während des Trainings mehrmals genutzt werden. Jeder Durchlauf heißt Epoche (*epoch*) [Aggarwal, 2015].

$$w^{t+1} = w^t + \Delta w \tag{80}$$

Es existieren verschiedene Algorithmen für die Änderung Δw. Viele Algorithmen nutzen die Gradienteninformation, d.h. nach jeder Änderung wird die Fehlerfunktion $\nabla E(w)$ wieder mit dem neuen Gewichtsvektor w^{t+1} berechnet [Bishop, 2006]. Einer der einfachsten Ansätze für die Anwendung der Gradienteninformation für die Auswahl der Gewichte in 80 besteht in der Bewegung auf einem Schritt in Richtung des negativen Gradienten, so dass [Bishop, 2006]:

$$w^{t+1} = w^t + \eta\Delta E(w^t) \tag{81}$$

Trotz der Einfachheit hat das einlagige Perzeptron eine wichtige Begrenzung: es kann nur auf linear trennbaren Daten angewendet werden [Minsky and Papert, 1969]. Reale Daten sind oft komplizierter und es ist notwendig mehrlagige Netze anzuwenden, welche fähig sind, das Problem zu lösen [Grossberg, 1973].

Mehrlagiges Perzeptron

Die Abbildung 12 (links) zeigt ein mehrlagiges Perzeptron (MLP), eines der am häufigsten

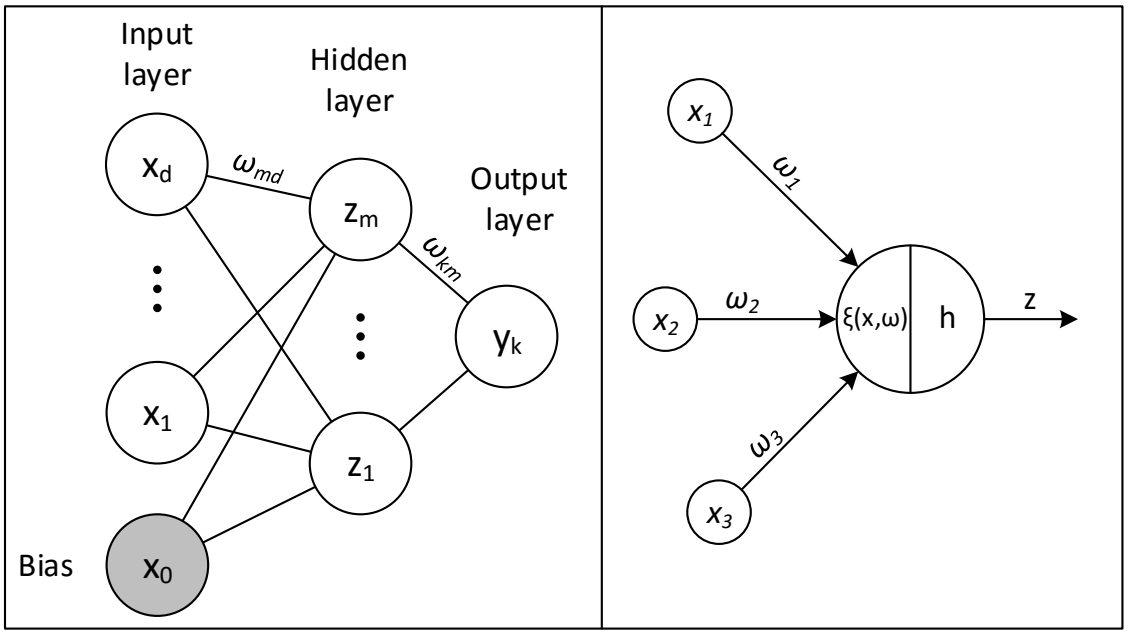

Abbildung 12: Mehrlagiges Perzeptron (*multilayer perceptron*)(links) und Model des künstlichen Neurons (rechts) (in Anlehnung an [Bishop, 2006])

verwendeten mehrlagigen neuronalen Netze. Dieses Netz besteht aus drei Schichten: einer Eingangsschicht (*input layer*), einer versteckten Schicht (*hidden layer*)(MLP kann auch mehrere versteckte Schichten enthalten) und einer Ausgangsschicht (*output layer*) [Kruse et al., 2015]. Es ist wert zu erwähnen, dass alle Neuronen einer Schicht mit Neuronen der nächsten Schicht vollständig verbunden sind. Ein solches Netz nennt man ein mehrlagiges Feedforward-Netz (*multilayer feed-forward network*). Die Topologie des Netzes wird automatisch definiert, wenn die Anzahl der Schichten und Anzahl der Neuronen in jeder Schicht bei dem Benutzer eingestellt wird [Aggarwal, 2015].

Das neuronale Netz nimmt die Eingänge $x = (x_1 \dots x_d)$ und überträgt diese durch einen Satz an Neuronen in der versteckten Schicht zur Ausgabe.

Im Fall des mehrlagigen Perzeptrons werden zuerst die Aktivierungen in der versteckten Schicht berechnet [Aggarwal, 2014, Bishop, 2006]:

$$net_j = \xi(x, w) = \sum_{i=1}^{D} w_{ji} x_i + w_{j0} \tag{82}$$

wobei $j = 1 \dots M$ die Indizes der Parameter und M die Anzahl der Neuronen in der versteckten Schicht des Netzes sind. Die Parameter w_{ji} sind die Gewichtswerte und w_{j0} der *Bias*.

Die Aktivierungen werden dann mittels einer nicht linearen Aktivierungsfunktion (*activation function*) $h(\cdot)$ in die Ausgabe des Neurons z_j verwandelt [Bishop, 2006]:

$$z_j = h(net_j) \tag{83}$$

Die Aktivierungsfunktion $h(\cdot)$ für Perzeptronen ist eine Sigmoidfunktion [Bishop, 2006]:

$$h(x) = \frac{1}{1 + \exp(-x)} \tag{84}$$

Folgend auf der Gleichung 82 können diese Werte wieder linear kombiniert werden, um die Aktivierungen in der Ausgangsschicht net_k zu berechnen [Bishop, 2006]:

$$net_k = \sum_{j=1}^{M} w_{kj} z_j + w_{k0} \tag{85}$$

wobei $k = 1 \dots K$ die Indizes der Parameter in der Ausgangsschicht des Netzes sind und K die gemeinsame Anzahl der Ausgänge ist. Diese Verwandlung entspricht der Ausgabeschicht des Netzes und w_{k0} dem Bias. Am Ende werden die Aktivierungen von Ausgangseinheiten mittels einer passenden Aktivierungsfunktion in die Ausgänge des Netzes y_k verwandelt [Bishop, 2006]:

$$y_k = h^{(2)}\left(\sum_{j=1}^{M} w_{kj} h^{(1)}\left(\sum_{i=1}^{D} w_{ji} x_i + w_{j0} \right) + w_{k0} \right) \tag{86}$$

wo $h^{(1)}$ und $h^{(2)}$ Aktivierungsfunktionen in der ersten (versteckten) Schicht und Ausgangsschicht entsprechend sind. Wie beim einlagigen Perzeptron können die Bias Parameter in den Satz der Gewichte aufgenommen werden [Bishop, 2006]:

$$y_k = h^{(2)}\left(\sum_{j=0}^{M} w_{kj} h^{(1)}\left(\sum_{i=0}^{D} w_{ji} x_i \right) \right) \tag{87}$$

Wie aus der Gleichung 87 ersichtlich enthält dieses neuronale Netz zwei Bearbeitungsschritte, von denen jeder einlagigen Perzeptron ähnlich ist, weshalb dieses Modell des neuronalen Netzes als mehrlagiges Perzeptron bekannt ist [Bishop, 2006]. Ein wichtiger Unterschied zum einlagigen Perzeptron besteht darin, dass das mehrlagige Perzeptron keine linearen Aktivierungsfunktionen in der versteckten Schicht nutzt. Deshalb stellt das Modell des neuronalen Netzes (Gleichung 87) eine nichtlineare Funktion von Eingängen x_i nach dem Satz der Ausgänge y_k dar, welche mittels der Parameter w einstellbar sind [Bishop, 2006].

Dieser Prozess kann als *forward propagation* der Information interpretiert werden .

Backpropagation (Fehler-Rückübertragung)

In einlagigen Perzeptronen ist der Trainingsprozess unkompliziert, weil der Ausgangswert mit dem Trainingslabel direkt verglichen werden kann. Die Optimierung wird mittels der Methode der kleinsten Quadraten durchgeführt. Die Anpassung der Gewichte wird in dem Fall mit einem Gradientenverfahren realisiert. Weil ein einlagiges Perzeptron nur ein einziges Neuron mit Gewichten als Ausgang nutzt, kann die Anpassung einfach implementiert werden. Im Fall der mehrlagigen Perzeptronen besteht das Problem darin, dass die *Ground Truth*

für eine versteckte Schicht nicht vorhanden ist, weil keine Trainingslabels mit Ausgänge von Neuronen in der Schicht assoziiert sind, sodass eine direkte Lösung wie im Fall von einlagigen Perzeptronen nicht möglich ist. Das notwendige Feedback kann man mittels des Fehler-Rückübertragungs-Algorithmus (*backpropagation*) bekommen [Aggarwal, 2015].

Der Fehler-Rückübertragungs-Algorithmus enthält zwei Phasen, welche während der Änderung der Gewichte für jeden Trainingspunkt durchgeführt werden:

1. *forward propagation*: diese Phase wurde oben beschrieben. Am Ende der Phase wird der Ausgangswert mit dem Klassenlabel verglichen und geprüft, ob das vorhergesagte Label falsch oder richtig ist.

2. *backpropagation*: Das Hauptziel in zweiter Phase ist Gewichte in die Rückwärtsrichtung anzulernen. Es wird mittels der Fehlerabschätzung von den Ausgängen in der bevorstehenden Schicht in Abhängigkeit von Fehlern in der nachstehenden Schicht realisiert. Die Fehlerabschätzung stellt eine Funktion von Fehlerabschätzungen und Gewichte der Neuronen in der vorherigen Schicht dar. Diese Information wird dann genutzt, um den Fehlergradient hinsichtlich der Gewichte von Neuronen zu berechnen und die Gewichte anzupassen.

Die Gleichungen für die Anpassung unterscheiden sich weniger von denen des einlagigen Perzeptron. Der Hauptunterschied besteht darin, dass nicht lineare Funktionen in den versteckten Schichten genutzt werden und es keinen direkten Vergleich zwischen den Ausgangswert und das Trainingslabel gibt, stattdessen wird die Fehlerabschätzung mittels des Fehler-Rückübertragungs-Algorithmus berechnet [Aggarwal, 2015].

Mehrlagige Perzeptronen sind stärker als *kernel SVM* in der Lage, die Trennungsfunktionen zu finden. MLP kann nicht nur eine Trennungsebene in einem multidimensionalen Merkmalsraum erfassen, sondern auch nicht zusammenhängende Klassenverteilungen mit verschiedenen Grenzen in unterschiedlichen Datenbereiche finden. Mit mehreren Neuronen und Schichten kann das MLP im Prinzip alle Funktionen approximieren bzw. die Daten klassifizieren. Trotzdem hat MLP folgende Nachteile [Aggarwal, 2015, Bishop, 2006]:

1. die Topologie des Netzes hat mehrere Kompromisse, welche die Analytik lösen muss. Eine größere Anzahl der Neuronen und Schichten bietet eine bessere Generalisierung, erhöht aber auch das Risiko für eine Überanpassung.

2. MLP braucht viel Zeit fürs Training und ist manchmal empfindlich für Rauschen.

Extreme Learning Machine als Neuronales Netz

Als eine Lösung für diese Probleme haben [Huang et al., 2004] eine neue Methode - *extreme learning machine (ELM)* vorgeschlagen. ELM stellt eine Gruppe von Methoden des maschinellen Lernens (einschließlich der einlagigen und mehrlagigen Perzeptronen) dar, in welchen die Neuronen in versteckten Schichten nicht angepasst werden müssen [Huang, 2015].

Diese Methode lässt einen SLP als einen Approximator arbeiten, welcher zufällig initialisierte Neuronen (zufällige Werte für die Eingangsgewichte für die Verbindungen zwischen Eingangs- und versteckten Schichten und zufällige Werte für Biases) in einer versteckten Schicht hat. Am Ende ist es notwendig, nur Ausgangsgewichte, welche die versteckte und Ausgangsschichten verbinden, analytisch zu berechnen anstatt sie anzulernen [Huang and Chen, 2007].

Es gibt drei Ebenen der Zufälligkeit in ELM [Huang, 2015]:

1. vollständig verbundene Neuronen, welche zufällig generierte Parameter haben

2. Verbindungen können zufällig erstellt werden. Nicht alle Eingangsneuronen sind mit einem bestimmten Neuron verbunden. Es ist möglich, dass ein lokaler Bereich mit einem versteckten Neuron assoziiert ist.

3. Ein verstecktes Neuron kann ein Subnetz darstellen, welches aus mehreren Neuronen besteht. Das bildet ein lokales, rezeptives Feld und führt zum Anlernen der lokalen Merkmale. Das bedeutet, dass einige lokale Teile von ELM mehrere versteckte Schichten enthalten können.

Als Aktivierungsfunktion in ELM können verschiedene Funktionen angewendet werden: Sigmoidfunktion, Radiale Basisfunktionen, Schwellwertfunktion, Fourier, quadratische Funktion, Wavelet-Funktion. Es ist möglich, sogar eine Kombination von diesen Funktionen zu nutzen.

ELM hat gute Ergebnisse für verschiedene Aufgaben [Nizar et al., 2008], [Sun et al., 2008], [Kasun et al., 2013] gezeigt. Im Vergleich zu MLP braucht ELM weniger Zeit für Berechnungen und kann als Aktivierungsfunktionen auch nicht differenzierbare Funktionen nutzen [Huang and Chen, 2007]. Dieser Algorithmus neigt auch weniger zur Überanpassung und braucht fast keine Anpassung der internen Parameter [Huang, 2015].

Deep-Learning-Methoden

Die Deep-Learning-Methoden basieren auf der Eigenschaft von neuronalen Netzen die interne Merkmalsdarstellung innerhalb der versteckten Schichten zu erstellen. Jede Schicht stellt einen Merkmalssatz für die nächste Schicht, damit wird die zusammengesetzte und ungesehene Datendarstellung erzeugt. Die größere Anzahl der Schichten mit einer größeren Anzahl der Neuronen in den Schichten ermöglicht es, stärkere Netze zu entwickeln, welche fähig sind, aussagekräftige Datendarstellungen zu produzieren [Aggarwal, 2014].

Das Training von *Deep Neural Networks* (DNN) ist eine komplexe Optimierungsaufgabe, weil die Anzahl der Hyperparameter bei einer großen Anzahl von Schichten und Neuronen riesig ist. Um die Aufgabe zu lösen, sind die neue Heuristik, die Vorkenntnisse über das Anwendungsgebiet und die Anwendung von einem oder mehreren Hochleistungsrechnern sowie eine große Datenmenge notwendig [Aggarwal, 2014].

Die drastische Vergrößerung der Datensatzgrößen und Rechenleistungen in den letzten Jahrzehnten haben eine breite Anwendung von DNN verursacht. Die frühesten DNN-Modelle im Bereich der Objekterkennung wurden nur auf kleine und ausgeschnittene Bilder angewendet. Die modernen Objekterkennungsmodelle sind fähig, hochauflösende Bilder zu bearbeiten, welche nicht unbedingt nur ein Erkennungsobjekt enthalten müssen [Krizhevsky et al., 2012], [Goodfellow et al., 2016].

Die Deep-Learning-Verfahren haben auch einen großen Einfluss gehabt und hohe Leistungen in Bereiche der Spracherkennung [Hinton et al., 2012], Fußgängerdetektion [Sermanet et al., 2013], Verkehrszeichenerkennung [Ciresan et al., 2012] und andere gebracht.

Trotz des großen Erfolgs in verschiedenen Bereichen haben die Deep-Learning-Verfahren auch folgende Nachteile: die Notwendigkeit einer großen Datenmenge und hoher Rechenleistungen zum Training. Das begrenzt die Anwendung von diesen Verfahren [Goodfellow et al., 2016].

2.9. Beurteilung des Klassifikators

Die Beurteilung der Qualität des Klassifikationsmodells hängt sehr stark von der Natur des Klassifikators ab. Sobald die Trainings- und Testdatensätze vorhanden sind, kann das Klassifikationsmodell aufgebaut werden. Es ist notwendig, die Genauigkeit des Models korrekt zu messen. Es führt zu zwei Herausforderungen [Aggarwal, 2014]:

- Methodologische Probleme. Diese Probleme stehen in Verbindung mit der Aufteilung der markierten Daten auf Trainings- und Testteile für die Beurteilung. Diese Auswahl beeinflusst direkt die Beurteilungsprozesse und kann zur Überschätzung oder Unterschätzung der Trennfähigkeit des Klassifikators führen. Es existieren verschiedene Ansätze, wie z.B. *holdout*, *bootstrap* und Cross-Validierung.

- Probleme der Quantifizierung. Diese Probleme sind mit einer numerischen Messung von der Qualität der Methode assoziiert, nachdem eine spezifische Methodologie für die Beurteilung gewählt wurde. Das kann z.B. die Genauigkeit, gewichtete Genauigkeit oder einen ROC(Receiver-Operating-Characteristic)-Wert sein. Andere numerische Messungen sind für Vergleich der relativen Leistung des Klassifikators entwickelt.

Validierungsschemen

Die Trainingsdaten sind für den Aufbau des Modells notwendig. Die anderen Daten - die Testdaten - werden angewendet, um die Leistung des Klassifikators zu bewerten.

Die ältesten Methoden haben den ganzen Datensatz sowohl für Training als auch für Test verwendet. Die Leistungen sind in dem Fall zu optimistisch, weil die Klassenlabels für das Modell schon bekannt sind. Es gibt bessere Methoden. Einige davon werden unten beschrieben.

Hold-out

Der Datensatz ist in zwei Sätze aufgeteilt - Trainingsdatensatz und Testdatensatz. Der Klassifikator lernt von den Trainingsdaten und seine Leistung wird dann auf den Testdaten gemessen. Der Anteil der Daten für das Training ist normalerweise 1/2 oder 2/3. Diese Methode hat ihre Grenzen. Nicht alle Daten werden für das Training genutzt. Je kleiner der Trainingsdatensatz, desto größer die Varianz des Modells. Wenn der Trainingsdatensatz zu groß ist, wird andererseits die geschätzte Genauigkeit (das Bias) zu klein. Das ist s.g. Bias-Varianz-Tradeoff. Es ist wichtig, zu erwähnen, dass beide Datensätze abhängig von einander sind, z.B. die Klassen, welche weniger in einem Satz präsentiert sind, werden stark präsentiert in dem anderen Datensatz [Dougherty, 2013].

Kreuzvalidierung

Bei diesem Ansatz wird der Datensatz nach einer zufällige Aufteilung in K in gleich große Teile aufgeteilt, Felder genannt, u.z. $\mathbf{D}_1, \mathbf{D}_2 \ldots \mathbf{D_K}$. Jedes Feld $\mathbf{D}_i$ wird als Testdatensatz betrachtet und der Trainingsdatensatz besteht aus den restlichen Feldern [Dougherty, 2013, Zaki and Meira Jr., 2014]

$$\mathbf{D} \setminus \mathbf{D}_i = \bigcup_{j \neq i} \mathbf{D}_j \tag{88}$$

Nach dem Training der Modells M_i auf dem Trainingsdatensatz $\mathbf{D} \setminus \mathbf{D}_i$ wird seine Leistung auf dem Testdatensatz $\mathbf{D}_i$ beurteilt, um die i-te Bewertung ψ_i zu bekommen. Dieser Prozess wird $\mathbf{K}$ Male wiederholt, sodass jeder Teildatensatz für den Test einmal genutzt wird [Dougherty, 2013]. Die erwartete Leistung des Klassifikators wird bei der Mittelwertbildung der Leistungen für alle Zyklen [Zaki and Meira Jr., 2014]:

$$\hat{\mu} = \frac{1}{\mathbf{K}} \sum_{i=1}^{\mathbf{K}} \psi_i \tag{89}$$

und die Varianz als [Zaki and Meira Jr., 2014]:

$$\hat{\sigma}^2 = \frac{1}{\mathbf{K}} \sum_{i=1}^{\mathbf{K}} (\psi_i - \hat{\mu}_\psi)^2 \tag{90}$$

Die Kreuzvalidierung kann mehrmals wiederholt werden, weil die initiale zufällige Aufteilung sicherstellt, dass die Felder jedes Mal unterschiedlich werden. Normalerweise wird das Parameter $\mathbf{K}$ gleich 5 oder 10 gewählt.

Ein Sonderfall der Kreuzvalidierung ist bei $\mathbf{K} = N$ (wo N Anzahl der Objekten im Datensatz ist), das ist der sogenannte *leave-one-out* Einsatz. Jeder Testdatensatz enthält in dem Fall nur ein Objekt und so viele Daten wie möglich werden für das Training genutzt. Dieser Ansatz ist sehr brauchbar für die Aufgaben, wo es wenig markierte Daten gibt, z.B. medizinische Diagnostik. Diese Methode ist rechenintensiv und führt zu einer hohen Varianz der Leistungen [Dougherty, 2013].

Bootstrap

Eine Alternative zu der Kreuzvalidierung stellt die *Bootstrap*-Methode dar, bei welcher der Datensatz mit einem Ersatz aufgebaut wird, d.h. das zuvor gewählte Objekt wird in den originalen Pool zurückgelegt und kann wieder gewählt werden, sodass die Kopien in einem Trainingsdatensatz existieren können. Das ist der beste Ansatz für das Resampling von sehr kleinen Datensätzen. Die übrigen Objekte werden für den Test genutzt [Dougherty, 2013]. Wegen der Probenentnahme wird die Wahrscheinlichkeit, dass ein gegebener Punkt gewählt wird, als $p = 1/n$ berechnet und die Wahrscheinlichkeit, dass er nicht ausgewählt wird, wird wie folgt berechnet [Zaki and Meira Jr., 2014]:

$$q = 1 - p = \left(1 - \frac{1}{n}\right) \tag{91}$$

Weil der Datensatz D_i n Punkte hat, kann die Wahrscheinlichkeit, dass das Objekt x_j nach n Versuche nicht ausgewählt wird, wie folgt berechnet werden [Zaki and Meira Jr., 2014]:

$$P(x_j \notin D_i) = q^n = \left(1 - \frac{1}{n}\right)^n \simeq e^{-1} = 0,368 \tag{92}$$

Das bedeutet, dass jede Bootstrap-Probe ungefähr 63,2% Punkte aus D enthält.

Im Vergleich zur Kreuzvalidierung vergrößert die Bootstrap-Methode die Varianz in jedem Feld. Das ist eine wünschenswerte Eigenschaft, weil es näher an den realen Experimenten ist [Dougherty, 2013]. Trotz dieses Vorteils erhält diese Methode eine zu optimistische Bewertung [Aggarwal, 2015].

Maße der Klassifikationsgüte

Verschiedene Maße der Klassifikationsgüte werden abhängig von der Art des Klassifikators angewendet [Aggarwal, 2014, Aggarwal, 2015]:

- In der Mehrheit der Klassifikatoren wird der Ausgang in Form eines Labels, welches mit dem entsprechenden Testobjekt assoziiert ist, vorhergesagt. Dieses Label wird mit dem *Ground-Truth*-Label verglichen. Danach werden die Klassifikationsleistungen berechnet.
- In anderen Fälle stellt der Ausgang einen numerischen Wert dar, welcher die Wahrscheinlichkeit für jedes Klassenlabel für das angegebenen Objekt zeigt. Es wird davon ausgegangen, dass die höheren Werte eine größere Likelihood darstellt, zu einer Klasse zu gehören

Wenn der Ausgang in Form von Klassenlabels dargestellt wird (siehe Tabelle 1), werden die *Ground-Truth*-Labels $gt(x)$ der Objekten x_i mit den vorhergesagten Labels $k(x)$ verglichen und können mittels Konfusionsmatrix (*confusion matrix*) (Tabelle 2) dargestellt werden [Aggarwal, 2014].

Tabelle 1: Beispiel der Klassifikation von einer Aufgabe mit mehreren Klassen

x	r(x)	k(x)
x_1	Klasse 1	Klasse 1
x_2	Klasse 1	Klasse 2
x_3	Klasse 2	Klasse 2
x_4	Klasse 1	Klasse 2
x_5	Klasse 2	Klasse 2
x_6	Klasse 2	Klasse 2
x_7	Klasse 2	Klasse 1
x_8	Klasse 1	Klasse 1

Tabelle 2: Konfusionsmatrix für die Aufgabe

		Tatsächliche Klasse	
		Klasse 1	Klasse 2
Klassifiziert als	Klasse 1	**rp** = 2	**fp** = 1
	Klasse 2	**fn** = 2	**rn** = 3

rp - Anzahl richtig positiver Klassifikationen

fp - Anzahl falsch positiver Klassifikationen

fn - Anzahl falsch negativer Klassifikationen

rn - Anzahl richtig negativer Klassifikationen

Auf der Basis vom Konfusionmatrix können die folgenden Werte berechnet werden [Aggarwal, 2015]:

1. Genauigkeit (*accuracy*). Das ist der Anteil von Testinstanzen, bei welchen der vorhergesagte Wert dem *Ground-Truth*-Wert entspricht und kann als Verhältnis von der Anzahl der richtig vorhergesagten Objekte (rp und rn) zur Gesamtzahl der Objekte berechnet werden [Aggarwal, 2014]:

$$Acc = \frac{rp + rn}{rp + rn + fp + fn} \tag{93}$$

2. Gewichtsensible Genauigkeit (*cost-sensitive accuracy*). In vielen Aufgaben haben einige Klassen nicht die gleiche Wichtigkeit wie andere. Dies ist sehr wichtig in unausgeglichenen Aufgaben, wie z.B. der Tumorerkennung. In dem Fall ist die falsche Klassifikation für bösartige Tumoren, welche eine kleinere Anzahl an Beispielen haben, viel schlechter als für die gutartigen. Es kann durch die Gewichte (*Cost*) $C_1 \dots C_k$ für die falsche Klassifikation von entsprechenden Klassen kontrolliert werden. Die gewichtete Genauigkeit kann wie folgt berechnet werden [Aggarwal, 2015]:

$$Acc = \frac{\sum_{i=1}^{k} C_i n_i Acc_i}{\sum_{i=1}^{k} C_i n_i} \tag{94}$$

wo $n_1 \dots n_k$ - Anzahl der Testinstanzen für jeder Klasse, $Acc_1 \dots Acc_k$ - die einzelne Genauigkeiten für jeder Klasse.

Die gewichtsensible Genauigkeit ist die gleiche ungewichtete normale Genauigkeit, wenn alle Gewichte $C_1 \dots C_k$ gleich sind [Aggarwal, 2015].

Es ist möglich, auch die andere Werte zu berechnen, welche die Klassifikationsleistungen beschreiben: Sensitivität (*sensitivity*), Spezifität (*specifity*), Präzision (*precision*), F-Wert und andere. Die Mehrheit davon ist für die binäre Klassifikation (zwei Klassen). Es gibt die Möglichkeit, die Werte auch für mehrklassige Aufgaben zu berechnen. Die Werte sind näher in [Aggarwal, 2014], [Aggarwal, 2015], [Zaki and Meira Jr., 2014] beschrieben.

2.10. Verfahren der Parameteroptimierung

Viele Verfahren des maschinellen Lernens haben interne Parameter, welche einen großen Einfluss auf die Leistungen haben können und abhängig von der Aufgabe angepasst werden müssen. Weiter werden die interne Parameter von den oben erwähnten Klassifikatoren und die Optimierungsstrategien zur Ermittlung der am besten geeigneten Werte für die Parameter beschrieben.

2.10.1. Relevante Parameter der Klassifikatoren

Die Tabelle 3 stellt eine Liste mit Klassifikatoren und ihren internen Parametern dar, welche in der Bibliothek für maschinelles Lernen "CARET" vorhanden sind.

Diese Parameter steuern direkt die Klassifikatoren und können zu drastischen Änderungen der Leistungen führen. Abhängig von der Aufgabe kann die Optimierung der Parameter viel Aufwand verursachen: die Tests des Klassifikators mit verschiedenen Parameterkombinationen auf einem großen Datensatz können viel Zeit und Ressourcen in Anspruch nehmen. Um das Problem zu lösen, wurden verschiedene Optimierungsstrategien entwickelt.

Tabelle 3: Parameter der Klassifikatoren

Klassifikator	Parameter Name	Beschreibung
Naive Bayes	fL	Faktor für die Laplace-Glättung
	useKernel	Parameter für die Aktivierung des Kerndichteschätzers zur Schätzung der Wahrscheinlichkeitsverteilung anstatt der Anwendung von der Normalverteilung
Logistische Regression	nIter	Anzahl der Boosting-Iterationen
k-Nearest Neighbors	k	Anzahl der betrachteten Nachbarinstanzen
Entscheidungsbaum	Conf	Schwellwert für Pruning
	M	Minimale Anzahl der Instanzen pro Blatt
Random Forest	mtry	Anzahl der zufällig ausgewählten Merkmale als Kandidaten bei jeder Trennung
	ntree	Anzahl der Bäume
SVM mit Kernel	cost (Linear, Radial, Polynom Kernel)	Kostenparameter, welcher die Glättung der approximierenden Funktion definiert
	sigma (Radial)	die invertierte Breite des Kernels
	degree (Polynomial)	Grad der polynomialen Kernelfunktion
	scale (Polynomial)	Maßstabparameter vom polynomialen Kernel für die Normierung von Mustern
Mehrlagiges Perzeptron	Size	Anzahl der Neuronen im versteckten Schicht(-e)

2.10.2. Optimierungsstrategien

Die Klassifikatorsparameter können manuell oder automatisch optimiert werden. Bei der manuellen Optimierung entscheidet der Entwickler, welche Parameterkombination für die Lösung der gegebenen Aufgabe besser geeignet ist. In dem Fall verlässt er sich auf eigene Kenntnisse oder auf das Expertenwissen, z. B. gibt es eine empfohlene Kombination der

Parameter für einige Klassifikatoren [Hsu et al., 2003]. Grid-Search und Random-Search-Verfahren repräsentieren die manuellen Optimierungsverfahren.

Grid-Search stellt ein Verfahren dar, welches eine erschöpfende Suche durch die Teilmenge der verschiedenen Kombinationen der Parameter durchführt. Die Parameterwerte können reellwertig und unbegrenzt sein, deswegen wird eine Diskretisierung und eine manuelle Einstellung der Wertgrenzen vor der Anwendung benötigt, wie z.B. die Parameter für den SVM-Klassifikator Kost (C) und Gamma(γ) auf der Abbildung 13 reelwertig sind und der Parameter Kost keinen oberen Grenzwert hat [Hsu et al., 2003]. Das Grid-Search-Verfahren leidet unter dem Fluch der Dimensionalität, aber die Berechnung kann einfach parallelisiert werden, weil die Berechnung der einzelnen Gitterpunkte unabhängig von einander ist.

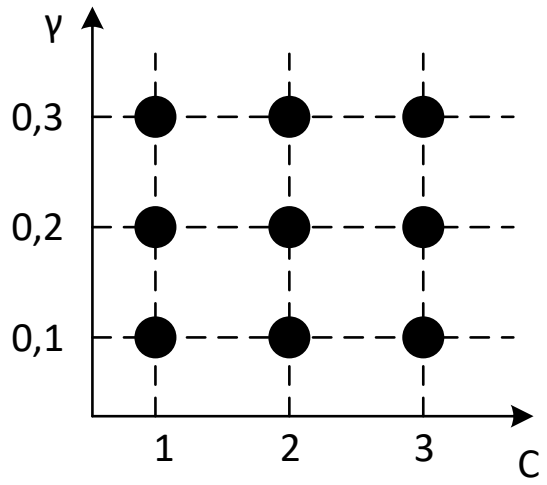

Abbildung 13: Grid-Search Beispiel

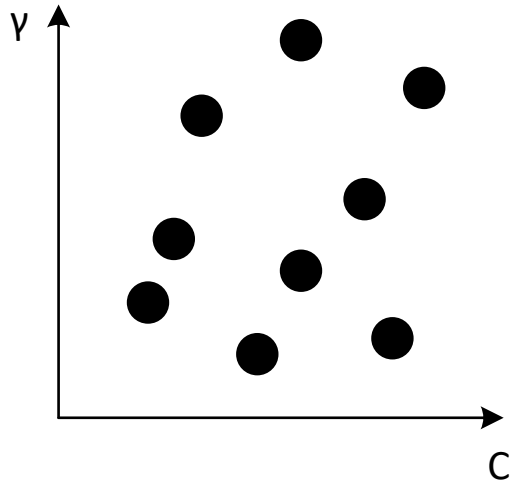

Abbildung 14: Random-Search Beispiel

Random-Search stellt eine Alternative für das Grid-Search-Verfahren dar. Die Kombinationen der Parameter werden in dem Fall zufällig aus einem begrenzten Parameterraum gewählt und dann berechnet (siehe Abbildung 14). In vielen Fällen lässt das Random-Search-Verfahren ähnliche oder bessere Leistungen im Vergleich zum Grid-Search erreichen [Bergstra and Bengio, 2012].

Die automatisierten Optimierungsstrategien sind mit bayesscher Optimierung, gradientenbasierter Optimierung und evolutionärer Optimierung repräsentiert.

Die bayessche Optimierung nutzt den Gaußprozess zur Modellierung der Aufgabe und die Bewertung der möglichen Verbesserung. Auf der Basis von bereits berechneten Kombinationen wurden die andere noch nicht berechnete Kombinationen bewertet und die mit der höchsten erwarteten Verbesserung wird weiter berechnet.

Bei der **gradienten-basierten Optimierung** wird der Gradient der Hyperparameter berechnet und dann werden mittels des Gradientenverfahrens die Parameter optimiert. Diese Optimierung ist sehr verbreitet für neuronale Netze und hat auch eine Anwendung für SVM [Chapelle et al., 2002] und logistische Regression [Do et al., 2007].

Die evolutionäre Optimierung wiederholt den Ansatz eines Merkmalsselektionsverfahren - eines evolutionären Algorithmus (siehe Kapitel 2.7). Es wird zuerst eine initiale Gruppe von zufälligen Kombinationen der Hyperparameter erstellt. Dann werden die Kombinationen berechnet und bewertet (z.B. mittels 10-Fach Kreuzvalidierungsverfahren). Die Kombina-

tionen werden auf der Basis der Leistungen sortiert - die schlechtesten werden durch die neuen Kombinationen ersetzt. Diese Schritte werden mehrmals wiederholt, bis eine genügende Leistung oder keine weitere Verbesserung erreicht wird.

3. Charakterisierung des spezifisches Schüttguttyps rezyklierte Gesteinskörnung

Abbildung 15: Gewaschene rezyklierte Gesteinskörnungen [Craven, 2010]

Bau- und Abbruchabfälle stellen Gemische aus mineralischen, metallischen und organischen Anteilen dar, welche eine entsprechend hochwertige Aufbereitung benötigen, um als rezyklierte Gesteinskörnungen wieder im Herstellungsprozess von Baustoffen verwendet werden zu können. Gesteinskörnungen stellen ein körniges Material dar, welches für die Betonherstellung geeignet ist. Man unterscheidet natürliche, synthetische (industriell hergestellte) und rezyklierte Gesteinskörnungen. Variierende Gehalte an porösen Partikeln mit hohem Ziegelanteil oder Zementsteingehalt, Kontamination durch organisches (Papier, Holz, Plastik) und anorganisches (Gips, Glas) Material erschweren die Sortierung der Gemische.

Die Bedeutung der Anwendung qualitativ hochwertiger Gesteinskörnungen in der Betonherstellung darf nicht unterschätzt werden. Zwischen 60% und 90% des Betonvolumens sind feine (Durchmesser $D < 4$ mm) und grobe (Durchmesser $D > 4$ mm) Gesteinskörnungen, welche damit einen großen Einfluss auf die Frisch- und Festbetoneigenschaften, die Mischungsanteile und die Verwendbarkeit des Betons haben [Nawy, 2008]. Wichtige Eigenschaften der Gesteinskörnungen für die Betonherstellung sind Porosität, Wasseraufnahme, Form, Größe, Druckfestigkeit und Anwesenheit schädlicher Substanzen [Mehta and Monteiro, 2005]. Besonders schädlich für die Betonherstellung sind Materialien, welche einen hohen Alkaligehalt haben. Alkalien sind Substanzen, welche mit Wasser Laugen (alkalische Lösungen) bilden [Mehta and Monteiro, 2005]. Die Substanzen haben einen starken Einfluss auf die Porosität, Dauerhaftigkeit und Druckfestigkeit des Betons. Glas und Gips haben einen hohen Alkaligehalt und müssen mit einer hohen Sicherheit aussortiert werden. Papier und organischer Müll verursachen Probleme mit der Vorbereitung und Härtung des Betons [Mehta and Monteiro, 2005].

3.1. Definition klassifizierender Bauschuttklassen auf Basis der DIN EN 12620 und DIN 4226-100

Bis 2008 galt DIN-Norm (Deutsche Industrie Norm) 4226-100 [DIN 4226-100, 2002] für Gesteinskörnungen in der Betonherstellung. Seitdem wurde die Norm durch den europäischen Standard DIN EN 12620 [DIN EN 12620, 2008] ersetzt. Inhaltlich basiert die neue Norm auf der alten. In DIN 4226-100 wurden vier verschiedene Liefertypen der Gesteinskörnungen definiert:

- Typ 1: Betonsplitt / Betonbrechsand
- Typ 2: Bauwerksplitt / Bauwerkbrechsand
- Typ 3: Mauerwerksplitt / Mauerwerkbrechsand
- Typ 4: Mischsplitt / Mischbrechsand

Die stoffliche Zusammensetzung der Liefertypen ist in der Tabelle 4 dargestellt.

Tabelle 4: Stoffliche Zusammensetzung der Liefertypen nach DIN 4226-100

<table>
<tr><th rowspan="2">Bestandteile</th><th colspan="4">Zusammensetzung
Massenanteil in Prozent</th></tr>
<tr><th>Typ 1</th><th>Typ 2</th><th>Typ 3</th><th>Typ 4</th></tr>
<tr><td>Beton und Gesteinskörnungen nach DIN 4226-1</td><td>≥ 90</td><td>≥ 70</td><td>≤ 20</td><td rowspan="3">≥ 80</td></tr>
<tr><td>Klinker, nicht porosierter Ziegel</td><td rowspan="2">≤ 10</td><td rowspan="2">≤ 30</td><td>≥ 80</td></tr>
<tr><td>Kalksandstein</td><td>≤ 5</td></tr>
<tr><td>Andere mineralische Bestandteile</td><td>≤ 2</td><td>≤ 3</td><td>≤ 5</td><td rowspan="2">≤ 2</td></tr>
<tr><td>Asphalt</td><td>≤ 1</td><td>≤ 1</td><td>≤ 1</td></tr>
<tr><td>Fremdbestandteile</td><td>$\leq 0{,}2$</td><td>$\leq 0{,}5$</td><td>$\leq 0{,}5$</td><td>≤ 2</td></tr>
</table>

Die neue Norm DIN EN 12620 gilt in Verbindung mit DIN 4226-101 [DIN 4226-101, 2017] und DIN 4226-102 [DIN 4226-102, 2017] für rezyklierte Gesteinskörnungen in Beton und die DIN EN 206-1 [DIN 206-1, 2001] / DIN 1045-2 [DIN 1045-2, 2008] gelten als die übergeordnete Normen für die Betonherstellung. Die DAfStb-Richtlinie "Beton nach DIN EN 206-1 und DIN 1045-2 mit rezyklierten Gesteinskörnungen nach DIN EN 12620" [DAfStb, 2010] regelt die Verwendung der Normen in der Betonherstellung.

In DIN 4226-101 wurden wie in der alten DIN 4226-100 vier Liefertypen definiert, obwohl sich die stoffliche Zusammensetzung voneinander unterscheidet - die Bauschuttmaterialien sind anders gruppiert und haben andere Bezeichnungen. Nach DIN EN 206-1/DIN 1045-2 dürfen in Beton nur die Liefertypen 1 und 2 verwendet werden (Tabelle 5).

Tabelle 5: Stoffliche Zusammensetzung der Liefertypen nach DIN 4226-101

	Bestandteile Massenanteil in Prozent				
Kategorie	$R_c + R_u$	R_b	R_a	$X + R_g$	FL
Typ 1	≥90	≤10	≤1	≤1	≤2
Typ 2	≥70	≤30	≤1	≤2	≤2

Rc: Beton, Betonprodukte, Mörtel, Mauersteine aus Beton

Ru: ungebundene Gesteinskörnung, Naturstein, hydr. geb. Gesteinskörnung

Rb: Mauersteine und Ziegel (nicht porosiert), Klinker, Steinzeug, Kalksandsteine, Mauer- und Dachziegel, Bimsbeton, nicht schwimmender Porenbeton

Ra: bitumenhaltige Materialien

Rg: Glas

X: sonstige Materialien (z. B. Ton und Boden, Metalle, Kunststoff Gummi, Gips)

FL: schwimmendes Material im Volumen

Im Vergleich zur alten Norm reduziert die neue Norm die Anforderungen zum zulässigen Anteil der Fremdbestandteile von 0,2% auf 1%. Die Prozentanteile von anderen Bestandteilen haben sich nicht geändert, obwohl nicht schwimmender Porenbeton in der alten Norm der Gruppe "Andere mineralische Bestandteile" mit zulässige Massenanteil ≤2% angehörte , was die höheren Anforderungen zur Kontrolle des Materials im Vergleich zur neuen Norm (wo nicht schwimmender Porenbeton der Gruppe R_b mit zulässigen Massenanteil ≤10% angehört) bedeutet.

Es ist wichtig, dass der DIN EN 12620 nicht genauer ausführt, zu wie vielen Anteilen Typ 1 aus Gesteins- oder Betonkörnung bestehen darf. Die Eigenschaften für ein Gemisch aus 90 M.-% Gesteinskörnungen + 10 M.-% der restlichen Kategorien sind ganz anders als für ein Gemisch aus 90 M.-% Betonkörnung + 10 M.-% der restlichen Kategorien. Dies führt zu Schwankungen bei der Herstellung von Recyclingbeton sowie in der Verarbeitung als auch in den Betoneigenschaften.

Tabelle 6: Mindesterkennungsraten für die Erfüllung der DIN-Normen

Bestandteile	Notwendige Einzelerkennungsraten für die Normen [%]	
	DIN EN 12620	DIN 4226-100
Beton und Gesteinskörnungen nach DIN 4226-1	≥ 90	≥ **90**
Klinker, nicht porosierter Ziegel	≥ 90	≥ **90***
Kalksandstein	≥ 90	≥ **90***
Andere mineralische Bestandteile	≥ 98	≥ **98**
Asphalt	≥ 99	≥ **99**
Fremdbestandteile	≥ 99	≥ **99,8**

* - ohne nicht schwimmender Porenbeton

Die Einzelerkennungsraten für die relevanten Bauschuttklassen müssen den Werten in der Tabelle 6 entsprechen, um die Normen DIN EN 12620 und DIN 4226-100 zu erfüllen. Die Anforderungen von DIN 4226-100 sind strenger, wie oben schon erwähnt.

3.2. Subklassen nach Norm

DIN 4226-100 hat 6 verschiedene Klassen definiert:

- Klasse 1 - Beton und rezyklierte Gesteinskörnungen
- Klasse 2 - nicht porosierter Ziegel und Klinker
- Klasse 3 - Kalksandstein
- Klasse 4 - andere mineralische Bestandteile (Leichtbeton, Porenbeton, porosierter Ziegel)
- Klasse 5 - Fremdbestandteile (Gips, Glas, Holz, Gipskarton)
- Klasse 6 - Asphalt

Die neue DIN EN 12620 legt 7 Klassen fest: Rc, Ru, Rb, Ra, Rg, X und FL (Erklärung für die Klassen sind in der Tabelle 5 dargestellt). Innerhalb jeder Klasse gibt es die Aufteilung in die Subklassen auf der Basis von chemischer Zusammensetzung und physikalischen Eigenschaften. In der Abbildung 16 ist die mineralogische Zusammensetzung von verschiedenen Betonproben dargestellt.

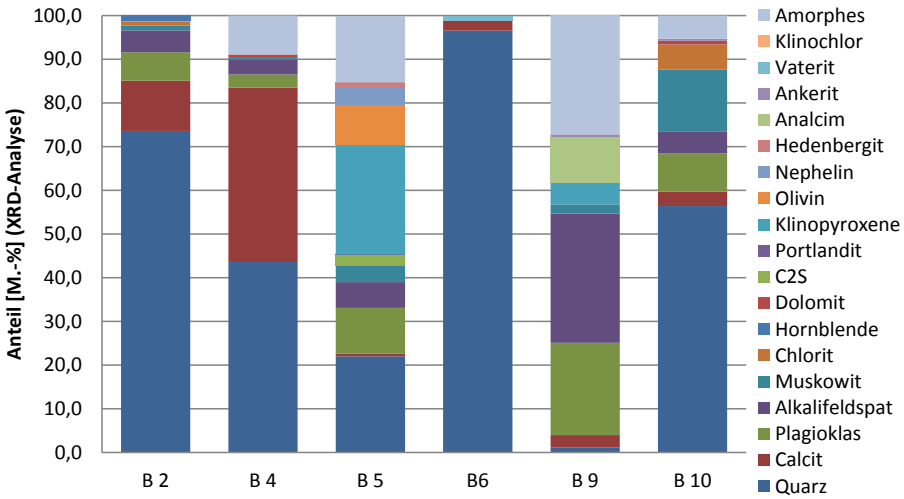

Abbildung 16: Mineralogische Zusammensetzung ausgewählter Betone [Linß, 2014]

Aus der Abbildung ist ersichtlich, dass die verschiedenen Proben von einem Material große Schwankungen in der Zusammensetzung haben.

Obwohl sich die Subklassen in stofflichen Zusammensetzung von einander unterscheiden, es ist schwierig, alle Subklassen zu erkennen, weil viele Komponenten nicht im VIS- und NIR-Bereich detektierbar sind. Es ist auch nicht sinnvoll, weil die DIN-Normen nur eine Aufteilung in viel grobere Klassen verlangen (siehe Tabelle 4 und 5).

4. Gerätetechnischer Entwicklungsstand im Bereich der Bauschutterkennung

Analyse- und Sortiergeräte im UV,VIS, NIR und Röntgenstrahlungsbereich

Die zurzeit auf dem Markt existierenden Sortiergeräte sind in der Tabelle 7 dargestellt. Die Analysegeräte im Bereich der Bauschutterkennung sind in der Tabelle 8 aufgelistet.

Tabelle 7: Analoga und Prototypen im Bereich der Bauschutterkennung

Entwickler, Hersteller	Produkt, Applikation	Einsatzbereich, Prüfgut	Technische Prinzip
Mogensen GmbH and Co. KG	MikroSort	Kalkstein, Kies, Marmor, Talkum, Barit, Quarz, Bauschutt (z. B.: Ziegelsplitt, Beton), Stein- und Meersalz	Anwendung verschiedener Sensoren, allein oder in Kombination: VIS-Kamera, NIR-Spektrometer, Röntgenstrahlung. Die Erkennung ist auf der Basis von Form oder Farbe/Helligkeit
RHEWUM GmbH	DataSort S	**Sortierung nach Helligkeit und Reflexion:** Marmor, Wolfram, Kohle, Quarz und Pyrit, Blei-Zinn-Erze, Talkum, Magnesium, Kalkstein, Kunststoffgranulat, Gips, auch Edelsteine: Rubine, Smaragde, Diamanten und Saphire. **Sortierung nach Echtfarben:** Feldspat, Gold (> 0,4 g/t), Kunststoffgranulat, Kunststoff-Flakes (PET, PE, PP, PVC), Recyclingglas (Flachglas und Hohlglas)	Anwendung der VIS-Kamera zur Erkennung auf der Basis von Farbe/Helligkeit bzw. Reflexion
Bühler GmbH	SORTEX (A, E, K, M, Z+)	Sortierung von Gemüse, Früchten, Beeren, Reis, Getreide, Samen, Gewürzen, getrocknetes Gemüse, Bohnen, Tee, Kaffee, Nüssen, Kunststoffen	Anwendung der VIS-Kamera oder SWIR-Kamera (InGaAs-Sensor) zur Erkennung auf der Basis von Helligkeit oder Farbe (abhängig vom Modell)
dataschalt Sortiertechnik GmbH	DataSort SISCO	Sortierung von Haselnüssen, Mandeln, Cornflakes, Erdnüssen, Kaffee, Tiefkühlprodukten, Mineralien (Bergbau)	

Entwickler, Hersteller		Produkt, Applikation	Einsatzbereich, Prüfgut	Technische Prinzip
S+S Separation and Sorting Technology GmbH		VARISORT C, FLAKE PURIFIER C, RAYCON BULK u.a.	Kunststoff, Glas, Holz, Metalle	Anwendung einer VIS-Kamera zur Erkennung auf der Basis von Farbe. **Raycon Bulk**: Anwendung von Röntgenstrahlung zur Detektion auf der Basis unterschiedlicher Dichte bzw. Absorptionseigenschaften
Entwickler: Fraunhofer-Institut für Optronik, Systemtechnik und Bildauswertung	Hersteller: Binder+Co AG	MINEXX	Sortierung von Quarz, Kalzit, Talkum, Kalkstein, Erze und Salze	Anwendung verschiedener Sensoren, allein oder in Kombination: VIS, NIR, UV-Sensoren zur Erkennung auf der Basis von Farbe/Helligkeit.
	ROC GmbH	ColorControl	Messung von Farbdifferenzen an Kunststoffgranulaten	Anwendung einer VIS-Kamera zur Erkennung auf der Basis von Farbe
		GranuControl	Sortierung von Kunststoffgranulaten nach Farbe und Formanalyse	Anwendung einer VIS-Kamera zur Erkennung auf der Basis von Farbe und Form
	OptoSort GmbH	Gemstar, BeltCompact	Sortierung von Gold, Platin, Diamanten, Kupfer, Nickel, Kalkstein, Feldspat, Magnesite, Talkum, Quarz	Anwendung einer VIS-Kamera zur Erkennung auf der Basis von Form und Farbe
TOMRA Sorting GmbH		PRO Primary COLOR	Sortierung von Kalkstein, Feldspat, Talkum, Steinsalz, Gold, Platinum	Anwendung einer Farbkamera, NIR-Kamera oder Röntgenstrahlung zur Erkennung auf der Basis von Farbe/Helligkeit, Form und Absorptionseigenschaften. **PRO Secondary LASER**: Anwendung eines Lasers zur Erkennung auf der Basis von Farbe, Form und Streueffekt.
		PRO Secondary and Tertiary COLOR, PRO Secondary LASER	Sortierung von Magnesit, Steinsalz, Talkum, Gold, Quarz, Phosphat, Borat, Feldspat, Branntkalk	
		PRO Granulate COLOR	Calcit, Quarz, Talkum, Borat, Steinsalz, Feldspat	
BEST (Belgian Electronic Sorting Technology) (Jetzt TOMRA Sorting GmbH)		Genius Optical Sorter	Kunststoffgranulat, Glas, Kunstharz	Anwendung der VIS- oder NIR-Kamera zur Erkennung auf der Basis von Farbe
LLA Instruments GmbH		KUSTA 2.2MSI	Gips, Kalksandstein	Anwendung einer Hyperspektral-Kamera zur Erkennung auf der Basis von charakteristischen Spektren der Materialien

Tabelle 8: Analyseverfahren

Entwickler, Hersteller	Produkt, Applikation	Einsatzbereich, Prüfgut	Technische Prinzip
HAVER & BOECKER	HAVER CPA 4	Bildanalyse von Korngrößen und Kornformen trockener und nicht agglomerierender Partikel von Schüttgütern. Verschiedene Bereiche von 30 μm bis 200 mm	Auswertung von Partikelgrößen und Partikelformen mittels einer CCD-Zeilenkamera
Retsch GmbH & Co.KG	CAMSIZER	Messung von Partikelgrößenverteilung und -form, sowie weiterer Parameter bei Pulvern und Granulaten. Bereich von 30 μm bis 30 mm mühelos	Auswertung von Partikelgrößen und Partikelformen mittels zweier CCD-Kameras
Micromeritics GmbH	Optisizer PSDA	Messung von Partikelgrößenverteilung und -form. Bereich von 75 μm bis 38,1 mm mühelos	Auswertung von Partikelformen und Partikelgrößenverteilung mittels einer CCD-Zeilenkamera
Microtrac	PartAn3D	Messung von Partikelgrößenverteilung und -form. Bereich von 15 μm bis 35 mm	Auswertung vom Größe, Form, Oberflächenrauhigkeit, Oberflächenspannung, Transparenz mittels einer Hochgeschwindigkeitskamera
Beckman Coulter GmbH	RapidVue	Messung von Partikelgrößenverteilung und -form. Analyse von Fasern, Stäben, Kristallen, Polymeren u.a. im Bereich von von 20 μm bis 2500 μm	Auswertung mittels einer CCD-Kamera
Malvern Instruments Ltd.	Sysmex FPIA3000	Messung von Partikelgrößenverteilung, -form und Durchsichtigkeit im Bereich von 0,8 μm bis 300 μm	Auswertung mittels einer CCD-Hochgeschwindigkeitskamera
Laboratoire Central de Ponts et Chausses (LCPC)	VDG 40	Bildanalyse von Korngrößen und Kornformen von Baumaterialien (Straßenbauhandwerk). Korngröße von 0,5 bis 50 mm	Auswertung mittels einer CCD-Zeilenkamera
WipWare Inc.	WipShape	Bestimmung des Längen-Breiten - Verhältnisses an Zuschlägen für Asphaltbeton	Anwendung von zwei CCD-Kameras zur Auswertung
University of Illinois	UI-AIA	Körngrößenverteilungen; Kornformparameter Kornvolumina, Kornoberflächen und –verteilungen	Auswertung mittels dreier Kamera: zwei VIS- und eine IR-Kamera

Entwickler, Hersteller	Produkt, Applikation	Einsatzbereich, Prüfgut	Technische Prinzip
Canty Process Technologie Chaussées	Solid Sizer	Verteilung von Durchmesser, Umfang, Fläche. Bildanalyse von Kornformen. Korngröße von 3 μm bis 3 mm	Auswertung mittels einer Schwarzweiß-Kamera

Die bereits existierenden Geräte sind in der Lage, einige Komponente von Bau- und Abbruchabfällen zu sortieren. Trotz großer Vielfältigkeit existiert momentan kein Sortier- oder Analysegerät, welches das komplexe Problem der Erkennung aller Bauschuttkomponenten nach DIN 4226-100 und DIN EN 12620 mit einer hohen Sicherheit (die Erkennungsrate von über 90% für z.B. Beton bis über 99.8% für Fremdbestandteile) lösen kann.

Das entwickelte System muss im Gegensatz zu den bereits auf dem Markt existierenden das Problem der Erkennung von Bauschuttkomponenten nach DIN 4226-100 lösen können. Die Klassen von Beton und Gesteinskörnungen, Klinker und nicht porosierten Ziegeln, Kalksandstein, sowie anderen mineralischen und Fremdbestandteilen müssen mittels des Systems mit einer hohen Erkennungsrate (von über 90% für z.B. Beton bis über 99,8% für Fremdbestandteile) erkannt werden. Dafür wird die Kombination von Bild- und Spektralinformationen genutzt, um die stark heterogenen Materialien auf der Basis fusionierter Information zu unterscheiden. Aus der Bildinformation werden nicht nur Form- und Farbmerkmale berechnet, sondern auch die Texturmerkmale, welche sehr relevant für die Erkennung von porösen Materialien sind.

5. Präzisierte Aufgabenstellung

Aus dem Gesamtziel der automatisierten Analyse von Bauschuttrezyklaten auf der Basis von Bild- und Spektralinformationen und den vorangegangenen Untersuchungen auf Bauschutt unter Verwendung von Farbbildern und maschinellen Lernverfahren [Linss et al., 2010], [Anding et al., 2011], [Linss et al., 2012a], [Linss et al., 2012b], [Garten et al., 2013], [Anding et al., 2013] folgen unterschiedliche, zu bearbeitende Teilaufgaben. Zuerst müssen alle notwendigen Bauschuttproben in geeignete Klassen zusammengefasst werden. Danach wird eine Bildaufnahme durchgeführt, um Objektbilder zu gewinnen und in Form eines Datensatzes zu ordnen. Die Bildaufnahme muss hierbei genügend Informationen über die Klassencharakteristik jedes Objektes zur Verfügung stellen, um eine hinreichende Trenngüte zu ermöglichen. Aus diesem Grunde sind im Rahmen dieser Arbeit auch Überlegungen und Untersuchungen zum notwendigen Wellenlängenbereich und der Sensorart anzustellen. Der gewonnene Datensatz muss dann bestimmten Kriterien entsprechen, wie z.B. der Objektanzahl pro Klasse und die Varietät der Klassen. Die Grundlage für die Klassenstrukturierung sind der Standard DIN 4226-100 und die Kompatibilität mit verwendeten Erkennungsalgorithmen. Deswegen besteht die Notwendigkeit, verschiedene Untersuchungen passender Klassen- und Merkmalsraumstrukturierungen durchzuführen.

Ein wichtiger Bestandteil der Arbeit ist die Auswahl und der Test am besten geeigneten Verfahren aus dem Bereich des maschinellen Lernens auf die Eignung zur Qualitätssicherung von Bauschuttrezyklaten. Das Hauptkriterium für die Bewertung ist die Erkennungssicherheit.

Die Bauschuttrezyklaten müssen mit einer hohen Sicherheit laut Standarten erkannt werden. Dafür gibt es die Notwendigkeit, folgende Teilproblemstellungen zu untersuchen und zu lösen:

- Vorüberlegungen und Untersuchungen zu Bild- und Spektrenaufnahme,
- Auswahl der geeigneten Merkmale,
- Klassifikatorwahl,
- Optimierung des Klassifikators für die Lösung des Problems.

Jede Teilaufgabenstellung enthält wiederum mehrere, zu lösende Unteraufgaben. So müssen die verschiedenen Wellenlängenbereiche vor dem Fokus der spektralen Charakteristik der verschiedenen Bauschuttklassen analysiert werden, für die Auswahl der geeigneten Merkmale müssen verschiedene Merkmalsselektions-/ Extraktionsverfahren geprüft werden. Das Zusammenspiel der Verfahren mit verschiedenen Klassifikatoren muss untersucht werden. Die Klassifikatorwahl stellt eine komplexe Aufgabe in Zusammenhang mit Merkmalsselektionsverfahren dar. Dann sollen die ausgewählten Klassifikatoren optimiert werden. Die am besten geeigneten Algorithmen werden dann gewählt. Somit kann eine automatisierte Klassifikation der einzelnen Objekte durchgeführt werden.

Das entwickelte System muss die folgende Klasse laut DIN 4226-100 mit angegebener Erkennungsrate unterscheiden, um die zuverlässige Erkennung der Bauschuttmaterialien zu gewährleisten:

Tabelle 9: Anforderungen an System und Verfahren

Bauschuttfraktion	Mindesterkennungsrate [%]
Beton und Gesteinskörnungen nach DIN 4226-1	$\geq$ 90,0
Klinker, nicht porosierter Ziegel	$\geq$ 90,0
Kalksandstein	$\geq$ 90,0
Andere mineralische Bestandteile	$\geq$ 98,0
Fremdbestandteile	$\geq$ 99,8

Das System muss am Ende der Analyse das Protokoll mit dem Prozentanteil oder mit der Stückanzahl von jeder Klasse erstellen.

Teil II.
Theorethische Untersuchungen

6. Analyse der Erkennungsaufgabe

Die Aufgabe der automatisierten Erkennung gehört zum Bereich des maschinellen Lernens. Das maschinelle Lernen stellt einen Satz von Methoden dar, welche automatisch Muster in Daten erkennen können und nutzen die erhaltene Information, um neue Daten vorherzusagen oder andere Entscheidungen unter Unsicherheit zu machen [Murphy, 2012].

Bei der automatisierten Erkennung werden die Testobjekte abgetastet, um die relevante Information zu bekommen. Die Information wird weiter durch Algorithmen bearbeitet und das Objekt wird der einen oder anderen Klasse zugeordnet. Die Abtastung kann mittels verschiedener Sensoren abhängig von der Aufgabe durchgeführt werden. Bei der Auswahl des Sensors muss man beachten, dass dieser die spezifische Objektinformation vollständig aufnehmen kann, um eine zuverlässige Klassifikation von Objekten auf der Basis von Information zu realisieren.

Die Klassifizierung stellt einen Prozess dar, bei welchem festgestellt wird, welcher Kategorie (Gruppe, Klasse) die neuen Instanzen angehören.

In dem Fall muss die Zusammensetzung einer Mischung festgestellt werden, was zum Teilgebiet des maschinellen Lernens - der automatisierten Klassifizierung gehört. Diesen Prozess führt der Klassifikator durch und ergibt am Ende die Vorhersage des Klassenlabels. Um das Klassenlabel richtig vorherzusagen, muss der Klassifikator zuerst antrainiert werden. Das Training wird unter Anwendung von Beispielen realisiert, welche markiert oder unmarkiert sein können. Im ersten Fall geht es um überwachtes Lernen (*supervised learning*), im zweiten um unüberwachtes Lernen (*unsupervised learning*). Die Markierung wird auf der Basis von Expertenwissen gemacht und so lässt sich eine bessere Klassifikationsleistung im Vergleich zu unüberwachten Lernverfahren erreichen [Duda et al., 2001], [Aggarwal, 2014], [Aggarwal, 2015]. Das heißt, die Bauschuttrezyklaten müssen zuerst gesammelt werden, dann ordnen Experten die Proben in bestimmten Klassen an und entwickeln auf der Basis von den Proben einen Datensatz, der dann zum Training und Test des Klassifikators genutzt wird. Die Lösung der Erkennungsaufgabe benötigt nicht nur einen Klassifikator. Eine Erkennungsroutine muss entwickelt werden, welche aus mehrere Schritten besteht. Dazu gehören die Merkmalsextraktion, die Merkmalsselektion, das Anlernen des Klassifikators und der Test des Klassifikators.

7. Auswahl der Algorithmen für Untersuchungen

Klassifikationsverfahren

Zurzeit existieren verschiedene Klassifikationsalgorithmen und diese Liste vergrößert sich jedes Jahr. In [Aggarwal, 2014] wurden alle Klassifikationsalgorithmen in 6 Gruppen zusammengefasst:

- statistische Klassifikatoren

- regelbasierte Klassifikatoren
- Entscheidungsbäume
- instanzbasierte Klassifikatoren
- Support-Vektor-Maschinen
- künstliche neuronale Netze

Alle Gruppen haben Vorteile und Nachteile, sowie spezifische Anwendungsbereiche. Aus der Vielfalt der Algorithmen wurden auf der Basis von Verfügbarkeit in den Bibliotheken für maschinelles Lernen, Leistungen in verschiedenen Aufgaben sowie auf der Basis des Erkenntnisstandes einige ausgewählt, welche verschiedene Klassifikationsansätze darstellen. Alle Untersuchungen wurden in Programmiersprache R implementiert und die Bibliotheken für Maschinelles Lernen "Caret" [Kuhn, 2009], "FSelector", "RWeka" wurden verwendet. Die Bibliothek "Caret" enthält zahlreiche Klassifikationsalgorithmen und einige davon den wurden für die Untersuchungen ausgewählt. Aus Gruppe der Statistischen Klassifikatoren wurden die Klassifikatoren Naive Bayes und Logistic Regression ausgewählt. Entscheidungsbäume sind mit C4.5-Algorithmus und Random Forest (ein Ensemble von einfacheren Entscheidungsbäumen) repräsentiert. k-Nearest Neighbor stellt die instanzbasierten Klassifikatoren dar. Die Gruppe von SVM-Klassifikatoren besteht aus SVM mit unterschiedlichen Kernen: RBF, Poly, Linear. Die künstlichen neuronalen Netze sind mit Multilayer Perzeptron und Extreme Learning Maschine (Ensemble und Batch-Verfahren) repräsentiert.

Merkmalsselektion und Merkmalsextraktion

Außer Klassifikationsalgorithmen wurden auch die Verfahren für Merkmalsselektion und Merkmalsextraktion ausgewählt. Weil die Merkmalsselektionsverfahren in Kombination mit Klassifikatoren getestet wurden, wurden nicht die Embedded-Methoden verwendet, da die Verfahren nur für eine begrenzte Anzahl von Klassifikatoren verfügbar sind.

Für die Merkmalsselektion wurde die Bibliothek "FSelector" verwendet. Die Bibliothek enthält verschiedene Algorithmen für die Merkmalsauswahl. Aus Filter-Verfahren wurden drei Algorithmen verwendet: Information Gain, chiSquare-Ranking und ReliefF-Filter. Die Wrapper-Verfahren wurden mit Simulated Annealing und genetische Algorithmen aus der Bibliothek "Caret" dargestellt. Die Merkmalsextraktionsverfahren wurden mit Hauptkomponentenanalyse und linearer Diskriminanzanalyse repräsentiert.

8. Realisierung der Erkennung von Bauschutt

8.1. Zusammenfassung in sinnvolle Oberklassen

Alle Subklassen können in größere Oberklassen zusammengefasst werden, um die Komplexität der Erkennungsaufgabe zu verringern und die Generalisierungsfähigkeit des Systems zu erhöhen. Die Zusammenfassung kann in kleinere Materialklassen erfolgen, was Klassen ergibt wie z.B. Beton, Kalksandstein, Leichtbeton u.a. Eine andere Möglichkeit ist es, die Subklassen nach DIN-Normen zusammenzufassen. Abhängig von der Norm erhält man 6 oder 7 Oberklassen. Im Endeffekt müssen alle Proben in Klassen nach DIN-Normen aufgeteilt werden, aber die Schwierigkeit besteht darin, dass die Subklassen innerhalb dieser Klassen eine hohe Variabilität haben (z.B. porosierter Ziegel und Leichtbeton in Klasse 4 nach DIN 4226-100 oder nicht porosierter Ziegel und Kalksandstein in Klasse Rb nach DIN EN 12620). Damit müssen die Untersuchungen bezüglich der Zusammenfassung in sinnvolle Oberklassen durchgeführt und die Anwendung der Aufteilung in Materialklassen und in Klassen nach DIN-Norm verglichen werden. Als DIN-Norm zum Vergleich wurde die Norm DIN 4226 untersucht, weil die Norm eine feinere Aufteilung hat und höhere Anforderungen an die Zusammensetzung der Liefertypen stellt.

8.2. Datensatzbereitstellung und Strukturierung

Die Proben wurden von Bauhaus Weimar bereitgestellt. Sie wurden zuerst manuell in Subklassen nach DIN 4226-100 sortiert. Insgesamt wurden 51 Pakete mit Proben geliefert: 10 verschiedene Betonproben, 1 Granitprobe, 10 Ziegelproben (porosiert und nicht porosiert), 8 Kalksandsteinproben, 8 Leichtbetonproben, 9 Porenbetonproben und 5 Gipsproben (als Ansetzgips, Gips und Gipskarton). Jedes Paket mit Probe enthält etwa 100 bis 500 Objekte, welche einer Subklasse angehören. Das Paket hat folgende Bezeichnung “Abkürzung vom Materialname + Probennummer”, z.B. “KS 1” für Kalksandstein. Das lässt später eine passende Zusammenfassung in Oberklassen testen.

8.3. Anforderungen an Bildaufnahme und Bilddatensatz

Um ein System zu automatisierter Erkennung von Bauschuttrezyklaten zu realisieren, muss zuerst dieses System antrainiert werden. Dafür wird ein Trainingsdatensatz benötigt. Dieser Datensatz muss genug Information erhalten, um die charakteristischen Probeneigenschaften widerzuspiegeln. Die Datensatzqualität beeinflusst verschiedene Faktoren:

- **Auflösung des Bildaufnahmesystems**
  Die Auflösung von Probenbildern muss hoch genug sein, um die Texturanalyse zu ermöglichen. Damit müssen die feinen Elemente der Oberflächen (die Mindeststrukturbreite) in den Farbbildern erkennbar sein.

- **Anzahl der Bildkanäle**
  Für eine zuverlässige Farbanalyse muss die Anzahl der Bildkanäle hoch sein.

- **Beleuchtung**
  Die Beleuchtung der Proben muss möglichst homogen sein und richtig angepasst werden,

um Unter- und Überbelichtung zu vermeiden. Die Beleuchtung darf keine Schatten produzieren, um die Erkennungsaufgabe nicht komplizierter zu machen.

- **Datensatzmarkierung**
  Die Proben müssen vor der Aufnahme richtig markiert sein. Die vorherige Sortierung und Markierung muss durch Expertenwissen realisiert werden. Eine falsche Anordnung führt zur Verschlechterung der Erkennungsroutine, welche auf dem Expertenwissen basiert.

- **Anzahl der Objekte pro Klasse**
  Die Anzahl der Objekte spielt eine große Rolle bei dem Training, da die relevanten Eigenschaften durch eine große Beispielanzahl angelernt werden können. Eine geringe Anzahl der Objekte führt zur Verschlechterung der Generalisierungsfähigkeit des Klassifikators.

8.4. Anforderungen an Spektrenaufnahme und Spektraldatensatz

Bei der Spektrenaufnahme müssen ähnliche Anforderungen wie bei der Bildaufnahme erfüllt werden, um die für die Erkennungsaufgabe relevante Information zu erhalten. Anstatt Bildauflösung und Anzahl der Bildkanäle werden spektrale Auflösung und Anzahl der Spektralkanäle wichtig. Der abgedeckte Spektralbereich muss die materialspezifische Wellenlänge umfassen. Die Bestandteile von Baumaterialien wie Carbonate, Hydroxide, Wasser haben die charakteristischen Absorptionsbanden im NIR-Bereich. Die Carbonate ($-CO_3$) haben charakteristische Absorptionsbanden bei 1850, 2000, 2350 und 2500 nm. Die Hydroxide (-OH) haben eine charakteristische Absorptionsbande in der Nähe von 1400 nm. Der Wasseranteil in Mineralien ist gut detektierbar bei 1450 und 1900 nm. Der Spektralbereich muss die Mehrheit der Wellenlängen umfassen. Die stabile und passende Beleuchtung hat auch einen großen Einfluss auf die Ergebnisse.

8.5. Anforderungen an den Hybriddatensatz

Im Hybriddatensatz sind die Informationen von Bilddatensatz und Spektraldatensatz verknüpft. Außer den oben genannten Anforderungen an die Bild- und Spektralaufnahme besteht die wichtigste Anforderung an den Hybriddatensatz darin, dass jedem aufgenommenen Bild sein jeweiliges spezifisches Spektrum zugeordnet werden muss, d.h. zu jedem Rezyklat-Objekt muss es einen Hybrid-Merkmalsvektor geben, der sowohl die Merkmale seines Farbbildes, als auch die Merkmale seines Spektrums enthält. Hierbei darf es keine Fehlzuordnungen von Spektren- zu Farbbildmerkmalen geben. Dies ermöglicht die Fusion der Informationen und eine zuverlässige Bewertung des Einflusses von verschiedenen Teilen (Bilder und Spektren) auf die Endergebnisse.

8.6. Vorüberlegungen zu geeigneten Bildmerkmalen

Auf der Basis von Probenbildern 17 ist es möglich, einige Vorüberlegungen zu geeigneten Merkmalen zu machen. Die Objekte haben ähnliche Formen und unterscheiden sich nicht so stark hinsichtlich ihrer Kontureigenschaften. Begründet ist dies darin, dass die Objektform

und Partikelgröße hauptsächlich durch den Zerkleinerungsprozess der Bauschuttabfälle im Backenbrecher entsteht und die Objektklassen somit keine großen Formunterschiede aufweisen. Aus diesem Grund weisen Formmerkmale eine geringere Relevanz für die Klassikationsaufgabe auf. Im Gegensatz haben die Farb- und Texturmerkmale größeres Potential für die Aufgaben, weil die Objekte unterschiedliche Farbe und Farbverteilungen haben (z.B. Gips, Beton und Ziegel) und die Oberflächen wegen der Porosität unterschiedlich aussehen. Die Merkmale wie Mittelwert, Abweichung des Grauwertes und Entropie können gute Aussagekraft für einige Klassen (wie Ziegel, Gips, Granit) haben. Anhand der Texturmerkmale wird es möglich, die porosierten von nicht porosierten Klassen zu unterscheiden.

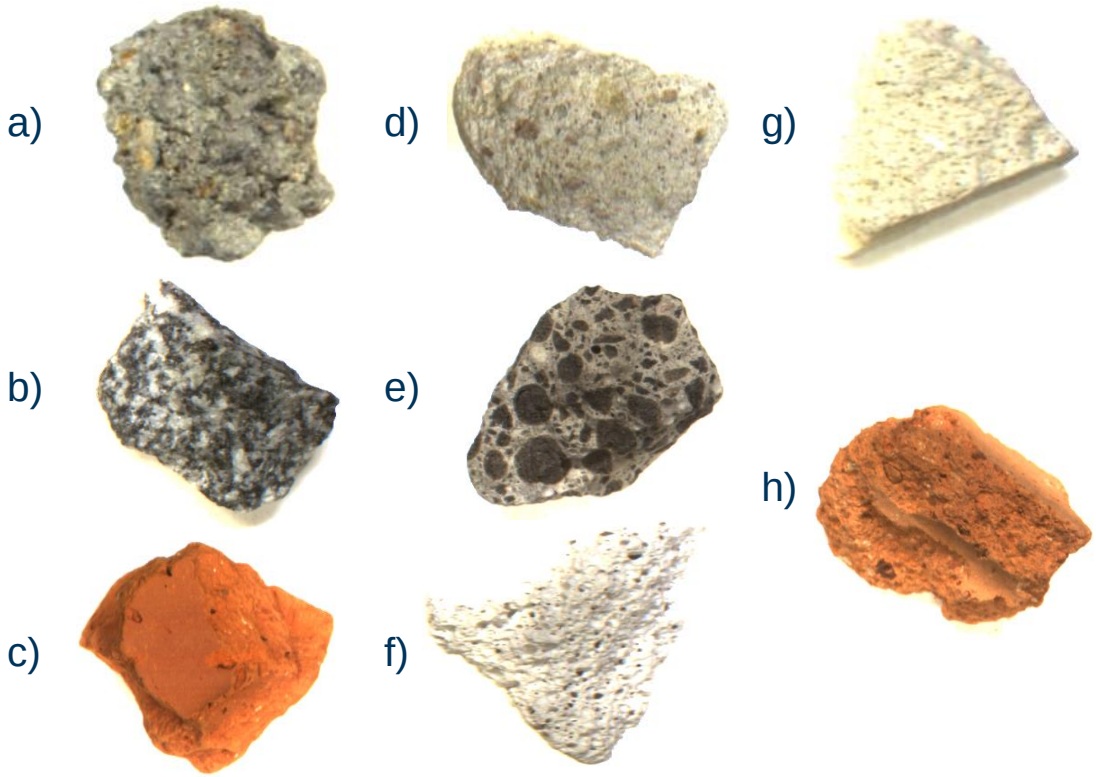

Abbildung 17: Bilder von verschiedenen Proben. a - Beton und Aggregate, b - Granit, c - Ziegel, d - Kalksandstein, e - Leichtbeton, f - Porenbeton, g - Gips, h - porosierte Ziegel

Die vorherigen Untersuchungen in dem Bereich [Anding et al., 2011] haben gezeigt, dass die Farb- und Texturmerkmale am geeignetsten für die Klassifikation von Bau- und Abbruchabfälle sind, was übereinstimmt mit diesen Vorüberlegungen.

8.7. Vorüberlegungen zu geeigneten Spektralmerkmalen

Die Spektralmerkmale stellen die Intensitätswerte bei den entsprechenden Wellenlängen aus dem untersuchten Spektralbereich dar. Weil das Spektrum eine kontinuierliche Kurve darstellt, unterscheiden sich die benachbarten Intensitätswerte nicht so stark von einander und deswegen ist die Korrelation zwischen benachbarten Wellenlängen hoch. Das führt zur Redundanz der Information. Materialspezifische Spektralmerkmale sind in dem Fall die

Absorptionsbände der Bestandteile des Materials. Das kann z.B. die Absorptionsbanden vom Wasser sein, welches in Gips eingebunden ist, bei 970, 1200, 1470, 1900 nm im NIR-Bereich [Irvine and Pollack, 1968].

Die kontinuierlichen Kurven von Spektren können mittels der Hauptkomponentenanalyse in Form komplexer Hauptkomponenten dargestellt werden, welche komplexe Merkmale sind.

8.8. Notwendigkeit der Merkmalsselektionsverfahren

Jeder der drei Datensätze (Farbbild-, Spektral- und Hybriddatensatz) enthält eine große Anzahl an Merkmalen, einige davon können nicht informativ (irrelevant) oder auch redundant sein. Um dieses Problem zu lösen, sind Merkmalsselektionverfahren notwendig. Die Anwendung dieser Methoden lässt nicht nur die relevante Information bewahren, sondern auch die wichtige dahinterliegende Abhängigkeiten zwischen Merkmalen und Objekteigenschaften finden. Das beeinflusst auch den Rechenaufwand und die Geschwindigkeit der Optimierungsprozesse positiv.

9. Spektrale Charakteristik von Bauschuttklassen

Bis zu diesem Zeitpunkt wurden sowohl in der Wissenschaft als auch in der Industrie umfangreiche Kenntnisse im Bereich der Spektroskopie gesammelt. In den Bereichen der Fernerkundung, der Mikroskopie sowie der Laboranalytik wurden in der Vergangenheit große Datenmengen in Form von Spektren verschiedener Wellenlängenbereiche aufgenommen und analysiert. Es entstand eine Vielzahl von Spektralbibliotheken, die zahlreiche Materialien und Wellenlängenbereiche umfassen.

Im Weiteren wurde die Online-Spektralbibliothek "ECOSTRESS" analysiert, welche die Spektren des Jet Propulsion Laboratory (JPL), der Johns Hopkins University (JHU) und des United States Geological Survey (USGS) umfasst und einen Spektralbereich von 0,4 bis 15,4 μm für verschiedene Materialien, wie Minerale, Gesteine, Böden, künstliche Materialien usw. abdeckt [Baldridge et al., 2009], [Hulley et al., 2017]. Die Informationen dieser Bibliothek erlauben einen Einblick in die Materialcharakteristika, welche zur Lösung der gegebenen Klassifikationsaufgabe geeignet sind.

JPL-Spektren

Die Mineralien wurden durch Zerkleinerung vorbereitet. Die zerkleinerten Proben wurden dann gesiebt, um die Probengröße 125−150 µm, 45−125 µm und <45 µm zu bekommen. Diese drei verschiedenen Probengrößen wurden genutzt, um den Einfluss der Größe auf den Reflexionsgrad zu messen. Die Spektren wurden in zwei Spektralbereichen angenommen: 0,4−2,5 µm und 2,0−15,4 µm. Das Spektralphotometer *Beckman UV5240* wurde für den Bereich 0,4−2,5 µm genutzt. Das Abtastintervall beträgt 0,001 µm für den Bereich von 0,4 bis 0,8 µm und 0,004 µm für den Bereich von 0,8 bis 2,5 µm.

Die direkte hemisphärische Reflexion wurde in diesem Spektralbereich mittels des Spektralphotometers *Perkin-Elmer Lambda 900 UV/VIS/NIR* gemessen. Die Spektren wurden mit dem Inkrement von 0,01 nm mit der Integrationszeit von 0,52 s (die Schrittgröße von 0,05 bis 5 nm) im UV-VIS-Bereich und mit dem Inkrement von 0,04 nm für 2,12 s (von 0,2 bis 20 nm) für den NIR-Bereich aufgenommen.

Die JPL-Spektralbibliothek enthält Spektren von 160 verschiedenen Mineralien, welche folgende Klassen bilden: Arsenate, Borate, Karbonate, Grundstoffe, Halide, Hydroxide, Oxide, Phosphate, Silikate, Sulfate, Sulfide und Wolframate.

JHU-Spektren

Die spektrale Bibliothek der Johns Hopkins University enthält Spekten von Mineralien und Meteoriten, welche im Bereich von 0,4−25 µm aufgenommen wurden.

Alle VIS/NIR (VNIR) Spektren wurden unter Anwendung des Beckman Instruments Model UV 5240 Spektralphotometers aufgenommen. Für die Aufnahme im Bereich 2,08−25 µm wurden zwei Nicolet FTIR Spektralphotometer verwendet.

USGS-Spektren

Die Spektren in der Bibliothek wurden mittels vier verschiedener Spektrometer erzielt:

1. Beckman 5270 im Bereich 0,2−3 µm
2. Analytical Spectral Devices (ASD) portabel Spektrometer im Bereich 0,35−2,5 µm
3. Nicolet Fourier Transform Infra-Red (FTIR) Interferometer-Spektrometer im Bereich 1,3−150 µm

4. NASA Airborne Visible/Infra-Red Imaging Spectrometer (AVIRIS) im Bereich 0,4−2,5 µm

9.1. Spektrale Charakteristik im VIS

Alle drei Bibliotheken enthalten einen großen Umfang an Mineralien und Baumaterialien. Aus der Vielzahl an spektralen Materialcharakteristiken wurden die für die Erkennungsaufgabe relevanten Materialklassen ausgewählt und analysiert.

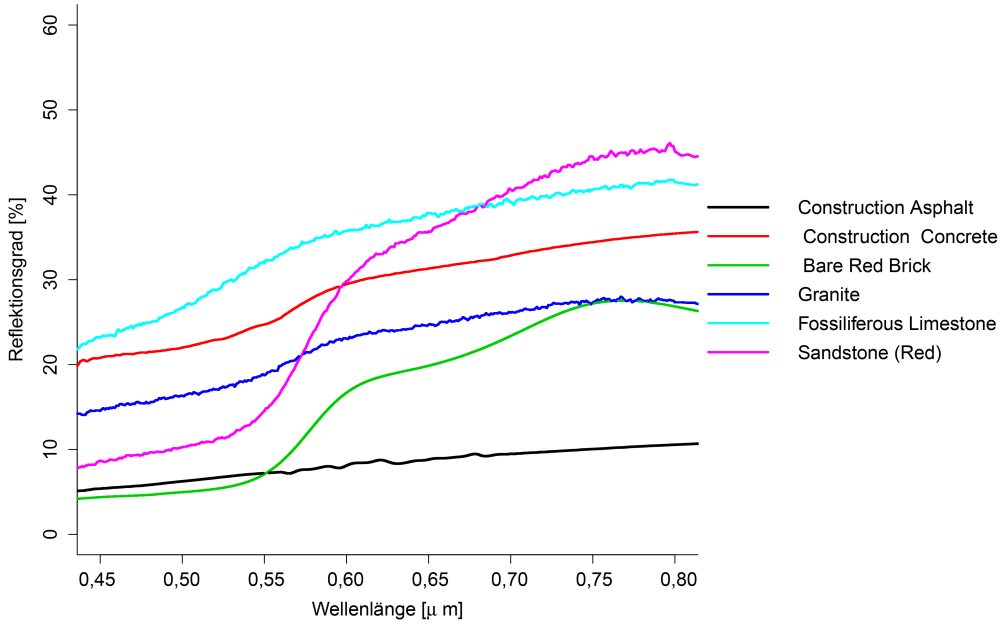

Abbildung 18: Spektren des Bauschutts im VIS-Bereich aus der JHU-Bibliothek

Aus der JHU-Bibliothek wurden folgende Exemplare für die Analyse ausgewählt: Asphalt (*construction asphalt*), Beton (*construction concrete*), Ziegel (*bare red brick*), Granit (*granite*), Kalkstein (*fossliliferous limestone*), roter Sandstein (*sandstone (red)*) (siehe Abbildung 18). Im VIS-Bereich zeigen die Spektren von Asphalt, Granit, Beton und Kalkstein Ähnlichkeiten. Der Unterschied zwischen diesen Spektralkurven liegt in der Ebene des Reflektionsgrades. Ziegel und Sandstein unterscheiden sich von anderen Klassen und untereinander mittels des Anstiegs des Reflektionsgrades im Bereich 0,55−0,75 µm. Die USGS-Spektralbibliothek enthält Spektren folgender Minerale: Gips (*gypsum*), Quarzit (*quartzite*), Dolomit (*dolomite*), Kalkstein (*limestone*) (siehe Abbildung 19). Quarzit und Kalkstein zeigen keine Reflektionsgradänderungen. Im Gegensatz dazu haben die Spektralkurven von Gips und Dolomit verschiedene Anstiegswinkel.

Die JPL-Bibliothek umfasst Spektren unterschiedlicher Mineralien und enthält außer den oben dargestellten Spektren auch noch weißen Marmor (*white marble*) (siehe Abbildung 20). Manche Materialien zeigen keine kontinuierlichen Spektren. Quarzit und Sandstein haben einen Anstieg des Reflektionsgrades bei 0,55 µm. Kalkstein und Dolomit zeigen kontinuierliche Anstiege. Andere Spektralkurven sind sehr ähnlich.

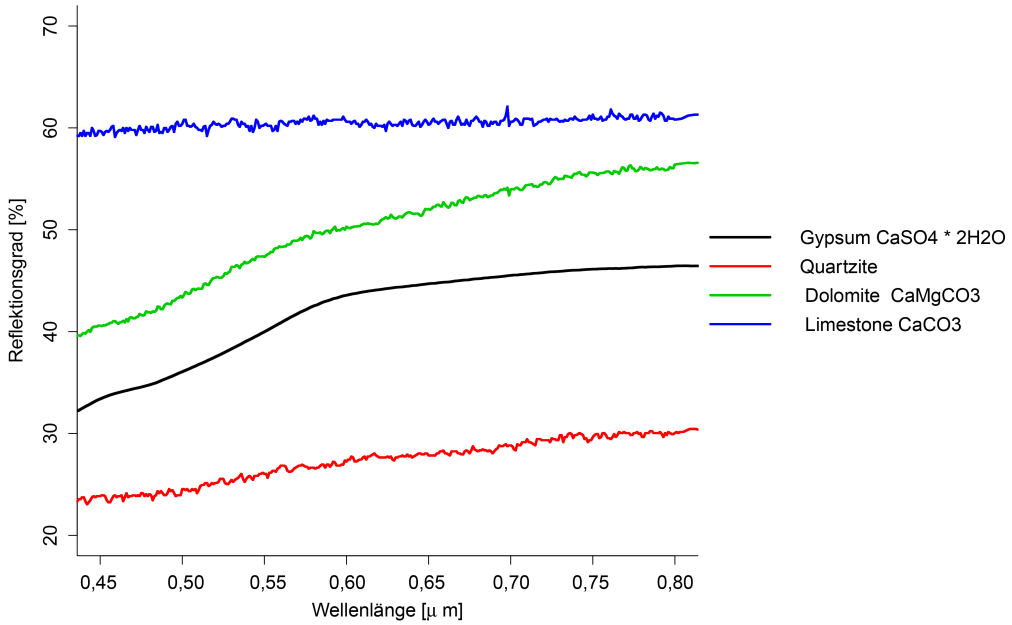

Abbildung 19: Spektren des Bauschutts im VIS-Bereich aus der USGS-Bibliothek

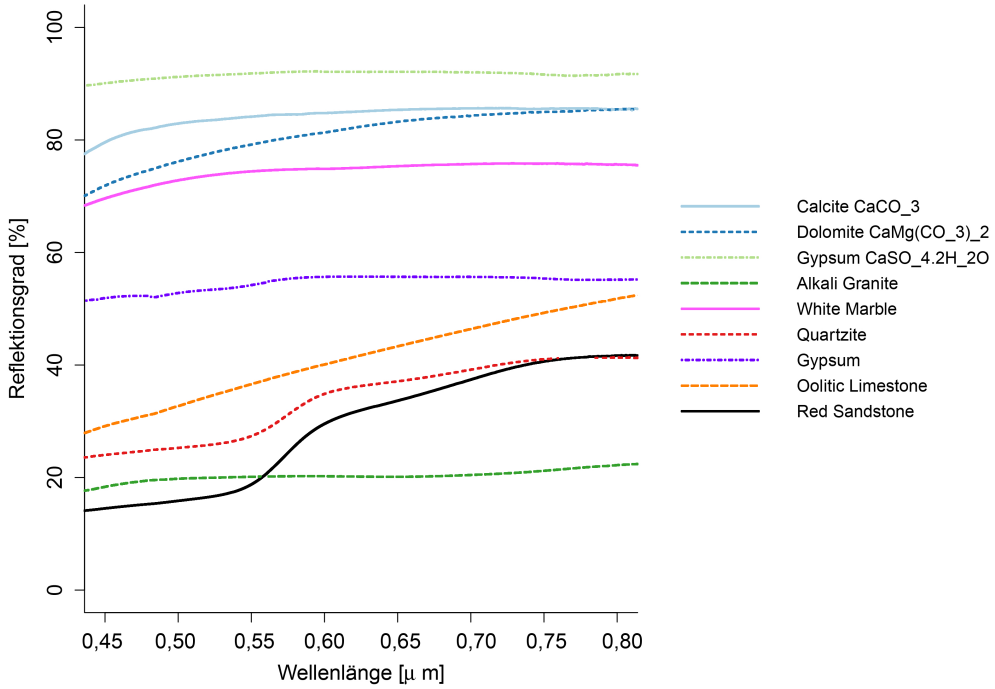

Abbildung 20: Spektren des Bauschutts im VIS-Bereich aus der JPL-Bibliothek

Auf der Basis dieser Spektren kann man schlussfolgern, dass es möglich ist, einige Materialien im VIS-Bereich zu unterscheiden. Brick, Sandstein (Bestandteil von Kalksandstein), Quarzit

(besteht zu 98% aus Quarz) und Dolomit haben einen erkennbaren "Fingerabdruck" im VIS-Spektralbereich.

9.2. Spektrale Charakteristik im NIR und IR

Im IR-Bereich zeigen die Proben der JHU-Bibliothek deutliche Unterschiede untereinander (siehe Abbildung 21). Asphalt zeigt im Gegensatz zu anderen Materialien eine kontinuierliche Spektralkurve. Granit ist ähnlich aber mit einer Absorptionsbande bei 1,9 µm. Ziegel hat einen gut erkennbaren Anstieg des Reflektionsgrades im Bereich 1−1,2 µm und eine Absorptionsbande bei 1,9 µm. Beton, Kalkstein und Sandstein haben mehrere Absorptionsbanden bei 1,4 µm, 1,9 µm, 2,3 µm und 2,5 µm. Man kann sie anhand der Anzahl an Absorptionsbanden und der Intensitätsänderung bei diesen Wellenlängen voneinander unterscheiden.

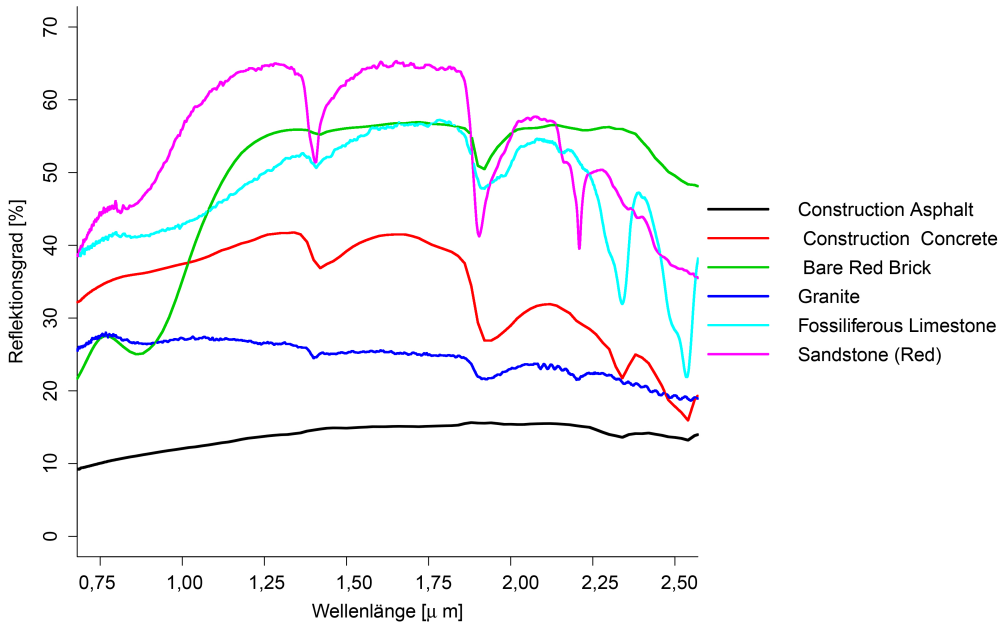

Abbildung 21: Spektren des Bauschutts im IR-Bereich aus der JHU-Bibliothek

Die Spektren von Dolomit und Kalkstein aus der USGS-Bibliothek sind sehr ähnlich und haben dieselben Absorptionsbänder bei 1,4 µm, 1,9 µm, 2,3 µm und 2,5 µm (siehe Abbildung 22). Quarzit zeigt keine Reaktion im IR-Bereich, Gips hat dagegen eine spezifische Spektralkurve und andere Absorptionsbänder im Vergleich zu Dolomit und Kalkstein bei 1,7 µm und 2,2 µm.

Die Spektren der JPL-Bibliothek weisen z.T. ein anderes Verhalten auf (siehe Abbildung 23). Hier unterscheiden sich Dolomit und Calcit nur in der Intensität des Reflektionsgrades im Bereich 2,0−2,2 µm. Außerdem haben sie Absorptionsbänder bei 1,9−2,0 µm und 2,3 µm. Die Spektralkurve von Gips hat neben mehreren Absorptionsbändern auch einen erkennbaren Abstieg im NIR. Die Proben von Granit und Quarzit zeigen schwache Änderungen im NIR-Spektralbereich. Marmor hat zwei charakteristische Absorptionsbänder bei 1,9 µm und 2,1 µm. Kalkstein hat weniger, dafür jedoch stärkere Absorptionsbänder.

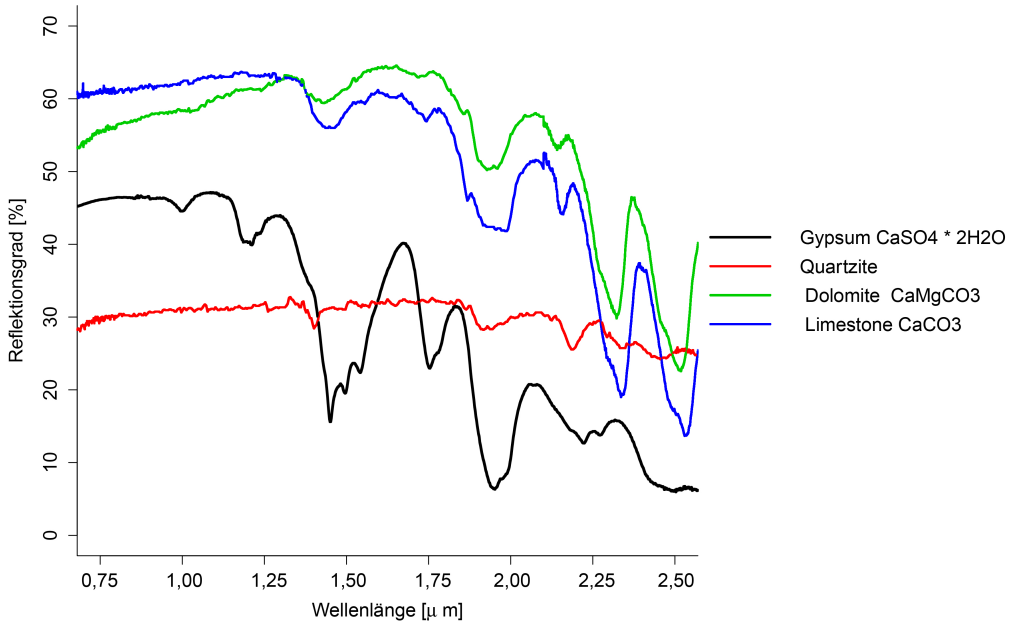

Abbildung 22: Spektren des Bauschutts im IR-Bereich aus der USGS-Bibliothek

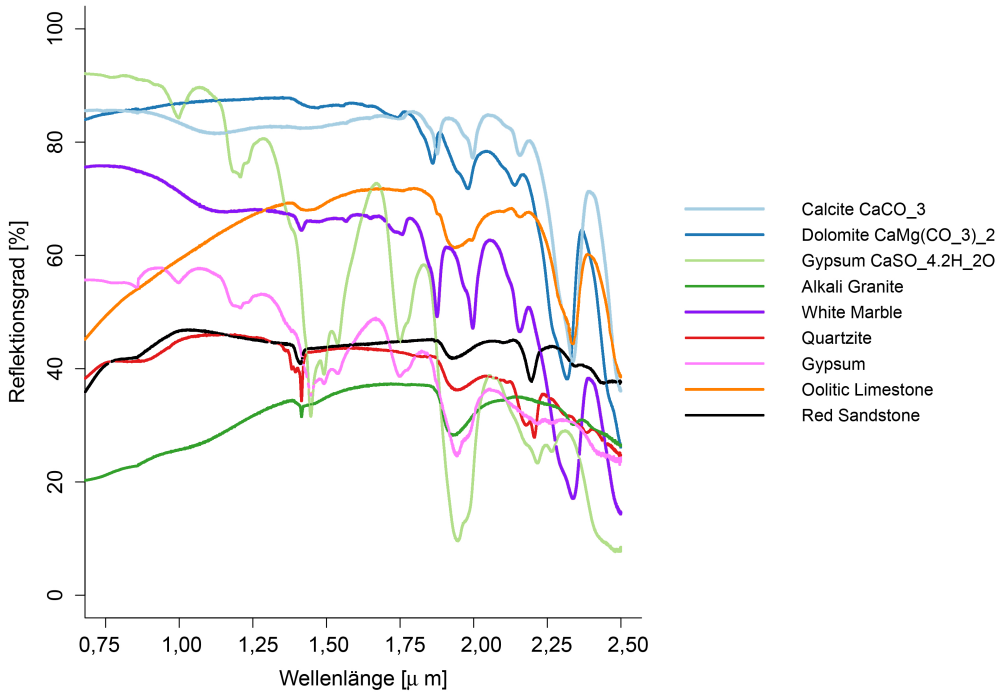

Abbildung 23: Spektren des Bauschutts im IR-Bereich aus der JPL-Bibliothek

Im NIR-Bereich gibt es deutlich mehr und gut erkennbare, spezifische Merkmale als im VIS, wie positiver oder negativer Anstieg als auch verschiedene Absorptionsbanden. Auf der Basis dieser Information kann man unterschiedliche Materialien und Minerale gut unterscheiden.

Für die weiteren Untersuchungen ist die Anwendung von einem NIR-Sensor, z.B. auf der Basis vom InGaAs-Detektor vielversprechend. Im NIR-Bereich befinden sich einige Absorptionsbanden (bei 1,4 µm, 1,9 µm, 2,3 µm und 2,5 µm), welche für die Charakterisierung von Baumaterialien verwendet werden können. Die üblichen ortsaufgelösten CCD-Sensoren können für die farbliche Charakterisierung im VIS-Bereich, sowie für die Form- und Texturanalyse angewendet werden, weil sie eine bessere räumliche Auflösung haben.

Teil III.
Experimentelle Untersuchungen

10. Bildaufnahme

10.1. Gerätetechnische Basis und applikationsspezifische Adaptation

Für die Bildaufnahme wurde der bereits entwickelte Aufbau des Fachgebietes Qualitätssicherung und industrielle Bildverarbeitung der TU Ilmenau verwendet. Der Aufbau wurde an der TU Ilmenau in Rahmen der Diplomarbeit von Herr Sebastian Dal-Canton für das Projekt "Autopetrographie" entwickelt [Dal-Canton, 2011] und besteht aus folgenden Teile (Abbildung 24):

- **Bildaufnahmeeinheit**: 3CCD-Zeilenkamera von Firma JAI und Zeiss-Objektiv
- **Beleuchtungseinheit** besteht aus zwei Hochleistungs-LED (light-emitting diode)-Zeilenbeleuchtungen für die Auflichtbeleuchtung und einer Hochleistungs-LED-Zeilenbeleuchtung als Durchlichtbeleuchtung unter dem Förderband,
- **Dosier- und Zuführeinrichtung** besteht aus zwei Förderbändern - geneigte und horizontale, einem Aufnahmebehälter und einem Auffangbehälter sowie einer Gruppe von Leitblechen zur Objektführung,
- **Gestell** aufgebaut aus fertigen Aluprofilen

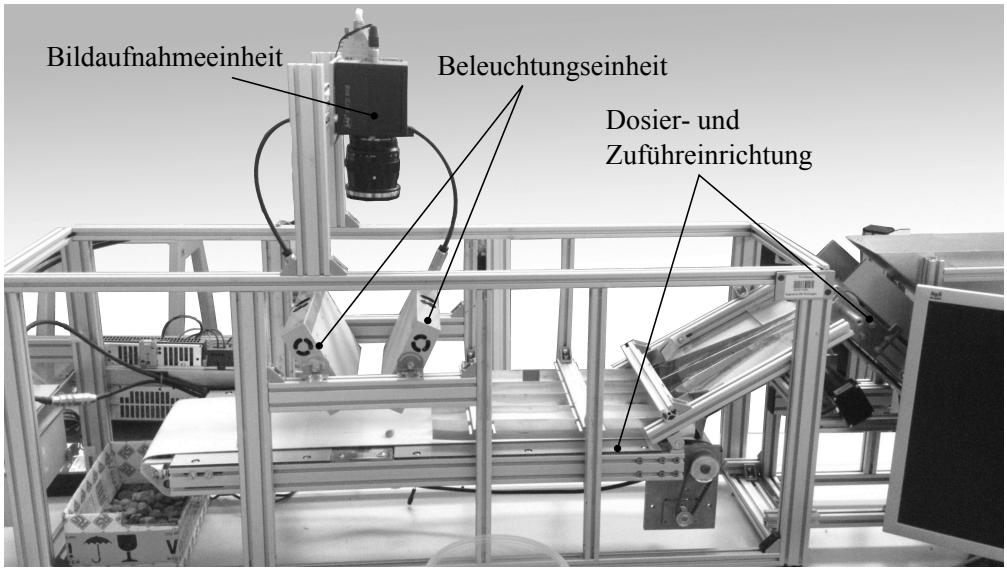

Abbildung 24: Bildaufnahmestand

Der Aufbau ist angepasst an die Bildaufnahme von Gesteinen und erfüllt aufgrund ähnlicher Objektgrößen ebenfalls die Anforderungen für die Bildaufnahme von Bauschuttrezyklaten in 8.3 [Dal-Canton, 2011].

10.2. Aufbau eines Gesamtdatensatzes von Bauschuttproben

Die Proben haben die Größe von 8 bis 16 mm und wurden vorher gesiebt und mechanisch gereinigt (ohne Spülung). Das Material wurde aus unbenutzten Baukomponenten durch Aufbruch und Zerkleinerung gewonnen. Die Datensatzstrukturierung und die Anzahl der Objekte pro Subklasse sind in der Tabelle 10 dargestellt.

Tabelle 10: Datensätze als Basis der Untersuchungen

Klasseneinteilung nach DIN 4226-100	Probenname	Materialklasse	Anzahl an Bsp-Objektbildern	
Beton- und Gesteinskörnung	B1	Beton	695	(5220)
	B2		687	
	B3		588	
	B4		549	
	B5		671	
	B6		833	
	B7		21	
	B8		570	
	B9		24	
	B10		582	
	Granit	Granit	745	(745)
Klinker / nicht porosierter Ziegel (Dichte > 2 g/cm^3)	Z4	Ziegel dicht	584	(1650)
	Z9		551	
	Z10		515	
Kalksandstein	KS1	Kalksandstein	594	(4808)
	KS2		541	
	KS3		561	
	KS4		771	
	KS5		644	
	KS6		526	
	KS7		535	
	KS8		636	
Andere mineralische Bestandteile	LB1	Leichtbeton	651	(5712)
	LB2		613	
	LB3		634	
	LB4		620	
	LB5		858	
	LB6		838	
	LB7		817	
	LB9		681	
	PB1	Porenbeton	608	(5080)
	PB2		532	
	PB3		631	
	PB4		599	
	PB5		582	

Klasseneinteilung nach DIN 4226-100	Probenname	Materialklasse	Anzahl an Bsp-Objektbildern	
	PB6		507	
	PB7		586	
	PB8		505	
	PB9		530	
	Z1		693	
	Z2		555	
	Z3		748	
	Z5	Ziegel porös	622	(4020)
	Z6		558	
	Z7		115	
	Z8		729	
	Ansetzgips		570	
	AnsetzgipsE2		529	
Fremdbestandteile	Gips	Gips	111	(1851)
	Gipskarton		376	
	Gipskarton1		265	

10.3. Datenaufteilung in Trainings-, Test- und Validierungspartitionen

Eine wesentliche Bedingung für die Anwendung überwachter maschineller Lernverfahren ist das Vorhandensein von sogenanntem Expertenwissen in Form der Ergebnisse der manuellen Zuordnung der Klassenlabel zu gegebenen Objekten. Dieses Apriori-Wissen muss als Input für das Training des Klassifikators genutzt werden. Während des Lernprozesses wird das angepasste Modell auf der Basis der vorgegebenen Merkmalsvektoren mit bekannter Klassenzuordnung entwickelt. Aus der gesamten Datenmenge wird ein Teil der Beispielobjekte (Lerndatensatz) fürs Training verwendet. Dieser Datensatz sollte ausreichend die Variabilität, die spezifischen Eigenschaften und wichtige Abhängigkeiten des gesamten Datensatzes aufweisen.

Für die Auswertung der Klassifikatoren wurde die 10-fache Kreuzvalidierung auf dem ganzen Datensatz genutzt.

11. Aufnahme klassenspezifischer Spektren

11.1. Entwicklung der Spektrenaufnahme

Im Kapitel 9 wurden die Spektren einiger Baumaterialien und Mineralien bewertet. Auf der Basis der theoretischen Untersuchungen wurde beschlossen, dass es notwendig ist, praktische Forschungen mit Baumaterialien im VIS- und IR-Bereich durchzuführen.

Es wurde ein Prüfstand eingerichtet, um VIS-Spektren aufzunehmen unter Verwendung des USB2000+ Spektrometers der Firma „Ocean Optics“, dessen Spektralbereich im Bereich von 200 bis 1100 nm mit einer Auflösung von 0,5 nm liegt. Die verwendete Lichtquelle ermöglicht eine stabile Beleuchtung im Bereich von 420 bis 770 nm. Es wurde ein Messkopf verwendet, der die lichtleitenden Fasern fixiert und anordnet. Die Bereiche 420-470 nm und 720-770 nm können wegen des geringen Signal-Rausch-Verhältnisses nicht für die Beurteilung der Spektren verwendet werden, welches zu Fehlern führen kann.

Für die Untersuchungen im NIR-Bereich wurde das Spektrometer PSS 2120 der Firma „Polytec“ mit InGaAs-Detektor und einem Spektralbereich von 1100 bis 2100 nm verwendet. Der Messkopf PSS-H-A03 mit integrierter Halogen-Lampe von der Firma „Polytec“ wurde für die Aufnahmen mit einem definierten festen Abstand angewendet. In einer weiteren Untersuchung werden die Ergebnisse der Spektrenaufnahme im IR-Bereich mit der Farbbildaufnahme vereinigt unter Verwendung des im Rahmen des Autopetrographie-Projektes entwickelten Bildanalysedemonstrators. Die Bildanalyse und Merkmalsvektorberechnung aus dem Farbbild wurde unter Verwendung eines Halcon-Skriptes von Frau Dr.-Ing. habil. Katharina Anding durchgeführt.

11.2. Aufbau eines Gesamtdatensatzes von Bauschuttproben

Auf der Basis der gelieferten Proben von der Bauhaus-Universität Weimar wurde die Spektrenaufnahme durchgeführt. Die Spektren wurden in Form von SPC- und TXT-Daten gespeichert. Das SPC-Format stellt eine blockweise-strukturierte Information über Spektren und Spektrenaufnahmebedingungen zur Verfügung. Die Daten wurden mit einem Programmskript in Matlab zusammengefasst und in eine ARFF-Datei umgewandelt, um sie in weiteren Untersuchungen zu benutzen.

Es wurden zwei Datensätze vorbereitet, jeweils einer für den jeweiligen Spektralbereich. Der VIS-Datensatz enthält 250 Proben und der IR-Datensatz 1041 Proben. Die Anzahl der Objekte pro Klasse hängt von der Anzahl der Subklassen pro Klasse ab. Für weitere Untersuchungen wurde ein weiterer Datensatz so strukturiert, dass er die Proben in die DIN-Oberklassen zusammenfasst (siehe Tabelle 11).

Tabelle 11: Datensatz für Untersuchungen im VIS- und IR-Bereich

Klasseneinteilung nach DIN 4226-100	Probenname	Materialklasse	Anzahl an VIS-Bsp-Objekt-Spektren	Anzahl an NIR-Bsp-Objekt-Spektren
Beton und Gesteinskörnung	B1, B2, B3, B4, B5, B6, B7, B8, B9, B10	Beton	50	155
	Granit	Granit	5	25
Klinker / nicht porosierter Ziegel (Dichte >2 g/cm^3)	Z4, Z9, Z10	Ziegel dicht	15	75
Kalksandstein	KS1, KS2, KS3, KS4, KS5, KS6, KS7, KS8	Kalksandstein	40	199
Andere mineralische Bestandteile	LB1, LB2, LB3, LB4, LB5, LB6, LB7, LB9	Leichtbeton	40	200
	PB1, PB2, PB3, PB4, PB5, PB6, PB7, PB8, PB9	Porenbeton	45	225
	Z1, Z2, Z3, Z5, Z6, Z7, Z8	Ziegel porös	35	105
Fremdbestandteile	Ansetzgips, AnsetzgipsE2, Gips, Gipskarton, Gipskarton1	Gips	25	57

11.3. Datenaufteilung in Trainings-, Test- und Validierungspartitionen

Bei den Spektren wurden wegen der geringeren Datensatzgröße die Untersuchungen durchgeführt, um ein passendes Validierungsschema zu finden. Die einfache Anwendung von 10-fach Kreuzvalidierung kann in dem Fall zur große Variation der Ergebnisse führen.

12. Anwendung des überwachten maschinellen Lernens auf den Bilddatensatz

Die Effektivität des Klassifikators hängt von mehreren Faktoren ab: von der Komplexität der Erkennungsaufgabe, der Ähnlichkeit der Klassen (Interklassenvariabilität) und der Streuung der Parameter innerhalb einer Klasse (Intraklassenvariabilität). Die Klassifikation bei stark unterschiedlichen Klassen kann mit einfachen und schnellen Klassifikatoren, wie z.B. dem Nearest Neighbor- oder J48-Klassifikator, gemacht werden. Im Falle ähnlicher Klassen steigt die Komplexität der Erkennungsaufgabe an und als Folge erhöhen sich die Anforderungen an den Klassifikator. Die Klassen sind in diesem Fall nicht linear trennbar, wodurch ein komplexer Klassifikator, wie z.B. eine SVM oder ein neuronales Netz notwendig wird.

Um einen passenden Klassifikator für die Aufgabe der Identifikation des Bauschutts festzulegen, wurden verschiedene Klassifikatoren auf der Basis verschiedener Datensätze getestet. Die Auswahl eines passenden Klassifikators hat große Bedeutung, weil sie die erreichbare Erkennungsrate definiert und demzufolge die Ergebnisse der Sortierung festlegt.

12.1. Implementierung der Merkmalsextraktion

Für die Extraktion der Bildmerkmale wurde die Bibliothek für maschinelles Sehen und Bildverarbeitung *MVTec HALCON* benutzt. Die Bibliothek enthält mehrere Operatoren, welche für die Berechnung von unterschiedlichen Merkmalen dienen. Unter Anwendung des Skriptes von Frau Dr. Anding wurden 235 Merkmale extrahiert. Die Merkmale stellen verschiedene Eigenschaften der Objekte dar wie Form, Farbe und Textur.

Einen Überblick über mögliche Bildmerkmale findet man in der Dissertation von Frau Dr.-Ing. habil. Anding [Anding, 2010].

Folgende Merkmale wurden für die Bildanalyse von Bauschuttrezyklaten berechnet:

- **Formmerkmale**: z.B. *Fläche*, *Flächenschwerpunkt*, *Kreisförmigkeit*, *Kompaktheit*, *Konturlänge*, *Konvexität*, *Flächendiameter*, *Anisometrie*, *Bulkiness*, *Struktur-Faktor*, *Radius* und *Orientierung der äquivalenten Ellipse*, *Radius des größten Innenkreises*, *verschiedene Momente*, *invariantes geometrisches Zentralmoment*, *Radius des kleinsten umschließenden Kreises*

Die Farbmerkmale wurden in jedem aus drei Kanälen separat berechnet:

- **Farbmerkmale**: *Energie*, *Korrelation*, *Homogenität*, *Kontrast*, *Entropie*, *Anisotropie*, *Mittelwert und Abweichung*, *Range*

Die Texturmerkmale wurden auf der Basis von Laws-Filter berechnet:

- **Texturmerkmale**: *Mittelwert und Abweichung für verschiedene Laws-Filter wie* ***ll, lr, ee, es, ew, se, ss, sw, rl, re, rs, rr, rw, wl, we, ws, wr***

12.2. Implementierung Merkmalsselektion

Der wichtigste Schritt für die Optimierung der Klassifikationsaufgabe ist die Auswahl der geeigneten Merkmale mittels Merkmalsselektionsverfahren. Nach der Merkmalsextraktion ist

es notwendig, die notwendige Anzahl an Merkmalen zu bestimmen. Diese müssen einen hohen Informationsgehalt und möglichst wenig Rauschen aufweisen. Im Gegensatz zur Merkmalsextraktion ist es notwendig, den optimalen Merkmalsatz nur einmal zu berechnen. Deshalb wurde dieser Schritt bei der Optimierung der Klassifikationsaufgabe durchgeführt. Danach wurde der berechnete Merkmalssatz aus den Daten extrahiert.

Für die Merkmalsselektion wurden verschiedene Algorithmen gewählt, welche den verschiedenen Prinzipien der Selektion entsprechen. So wurden Filterverfahren, Wrapper- und Embedded-Methoden angewendet und ihre Leistungen miteinander verglichen, um eine optimale Lösung zu finden. Aus den Filterverfahren wurden der ChiSquare-, InfoGain- und ReliefF-Filter gewählt.

Für die Bildanalyse wurde die Merkmalsselektion nach der Merkmalsextraktion (mittels HALCON) durchgeführt.

12.3. Evaluierung geeigneter Klassifikationsverfahren in R

Um die Erkennungsaufgabe zu lösen, wurden die Klassifikatoren aus der Bibliothek des Maschinellen Lernens - *caret* (Programmiersprache R) getestet. Folgende Klassifikationsalgorithmen wurden zuerst auf dem Datensatz aus 51 Subklassen (siehe Tabelle 10) angewandt:

- *C4.5 Tree (J48)*
- *Naive Bayes (NB)*
- *k-Nearest Neighbors (kNN)*
- *Random Forest (RF)*
- *Logistic Regression (LogitBoost)*
- *Support Vector Machine mit linear Kernel (svmLinear)*
- *Support Vector Machine mit polynominal Kernel (svmPoly)*
- *Support Vector Machine mit Gaussian (RBF) Kernel (svmRadialSigma)*
- *Multilayer Perceptron (MLP)*
- *Extreme Learning Machine (ELM)* (Neuronales Netz mit *Boosting* Elementen)

Die Ergebnisse (Abbildung 25) zeigen, dass die Klassifikatoren *Random Forest*, *LogitBoost*, SVM (alle drei verschiedene Kernen) und MLP höhere Erkennungsraten haben. Die Klassifikatoren *C4.5*, *k-Nearest Neighbors*, *Naive Bayes* zeigen eine niedrige Erkennungsrate (im Vergleich zu den bereits genannten Klassifikatoren) wegen der hohen Komplexität der Erkennungsaufgabe. Die Klassen sind linear schwer trennbar für solchen Klassifikatoren. Eine sehr hohe Klassenanzahl mit einer geringen Interklassenvariabilität stellt hier die Schwierigkeit der Klassifikationsaufgabe dar. Es ist jedoch wenig sinnvoll alle Subklassen zu unterscheiden, da die gestellte Aufgabe die Analyse und spätere Sortierung der Proben in die nach geltender DIN definierten Klassen ist und somit eine tiefe Trennung in alle 51 Subklassen keinen Gewinn mit sich bringt. Deshalb ist es rational die Subklassen in größere Gruppen zusammenzufassen

und damit die Interklassenvariabilität zu erhöhen, als auch die Klassifikationsaufgabe zu optimieren. Es gibt zwei mögliche Lösungen für die Zusammenfassung: in Oberklassen laut DIN-Norm oder in Obergruppen auf Basis des gegebenen Materials. Für den ersten Fall wurden die Klassen von 1 bis 6 verwendet. Im zweiten Fall wurden 8 Gruppen (Beton, Gips, Granit, Kalksandstein, Leichtbeton, Porenbeton, porosierter und nicht porosierter Ziegel) verwendet.

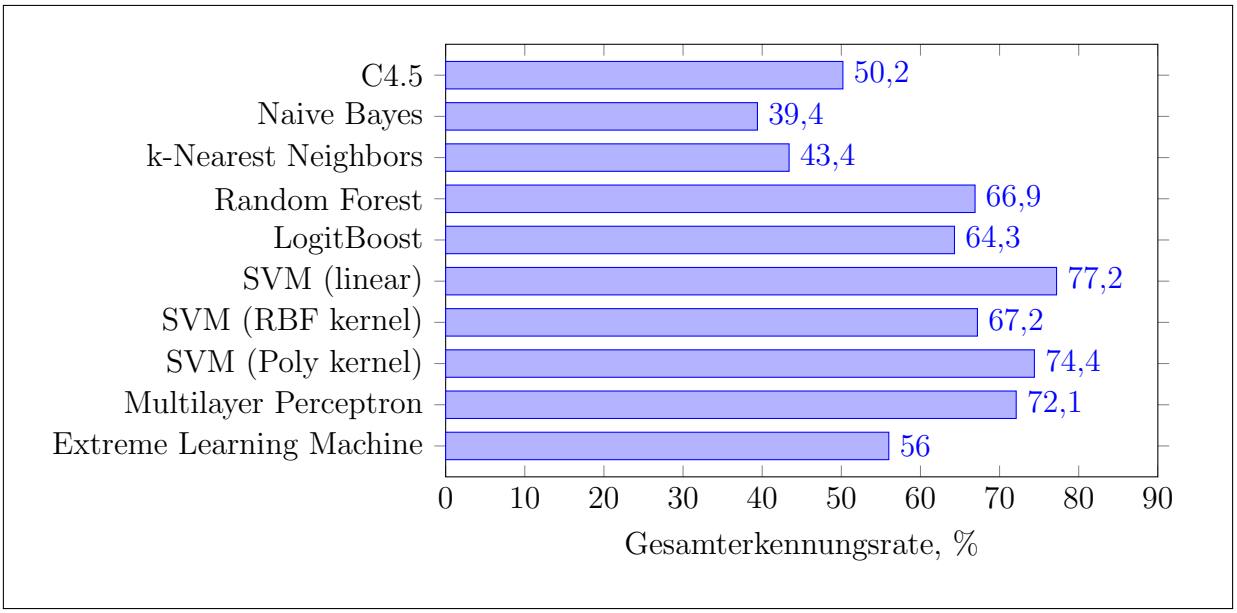

Abbildung 25: Visualisierung der Leistung ausgewählter Klassifikatoren auf dem Datensatz mit 51 Subklassen

Weitere Untersuchungen wurden mit zwei Datensätzen gemacht, den Oberklassen laut DIN-Normen und den Materialklassen. Es wurde auch die Anzahl der Klassifikatoren auf die am besten geeigneten begrenzt - SVM, MLP, RF, Logistische Regression. Die Ergebnisse der Klassifikation der 51 Subklassen können in einem ersten Schritt künstlich in zusammengefasste Oberklassen bzw. Materialklassen berechnet werden. Tabelle 12 zeigt die Resultate dieser Umformung in Oberklassen und Tabelle 13 in Materialklassen. In den Tabellen sind die Ergebnisse in Form von Einzelerkennungsraten (EER) und Gesamterkennungsraten (GER) dargestellt. Der Unterschied zwischen der künstlichen Zusammenfassung vor und nach dem Klassifikationsprozess liegt in der Generalisierung. Bei 51 Subklassen zielt der Klassifikator auf eine maximale Trennung zwischen den Subklassen ab, was nicht optimal im Sinne sinnvoller Oberklassen ist, weil die Grenzen im Merkmalsraum sehr eng definiert werden und das zu einer schlechten Generalisierung führt.

Obwohl die Ergebnisse gut sind (manche Klassifikatoren zeigen die Gesamt-Erkennungsrate über 90%), ist es rationeller die Zusammenfassung in größere Gruppen vor dem Klassifikationsprozess durchzuführen, damit die Generalisierung erhöht und der Rechenaufwand verringert wird. Die SVM-Klassifikatoren können jetzt schneller optimiert werden, weil die Rechenzeit quadratisch von der Anzahl der Klassen abhängt.

Tabelle 12: Erkennungsrate (Untersuchung mit 5 künstlichen Oberklassen)

	C4.5		NBayes		kNNearest		LogitBoost		SVM (RBF)		SVM (linear)		RF		ELM		svmPoly		MLP	
	EER [%]	GER [%]	EER [%]	GER [%]	EER [%]	GER [%]	EER [%]	GER [%]	EER [%]	GER [%]	EER [%]	GER [%]	EER [%]	GER [%]	EER [%]	GER [%]	EER [%]	GER [%]	EER [%]	GER [%]
Klasse 1	78,9		67,1		82,1		87,4		93,8		94,6		91,8		83,7		94,8		94,0	
Klasse 2	89,8		77,2		73,6		92,2		93,5		97,2		85,2		74,0		89,5		87,7	
Klasse 3	82,6	85,8	76,1	76,2	75,3	81,2	84,2	89,6	93,8	94,2	95,5	96,2	90,8	92,0	96,2	84,4	95,1	95,4	93,9	93,7
Klasse 4	90,3		80,4		83,4		91,9		95,3		97,2		94,0		83,2		96,7		94,9	
Klasse 5	79,1		72,4		82,6		84,3		88,5		93,4		87,5		86,8		93,3		88,2	

Tabelle 13: Erkennungsrate (Untersuchung mit 8 künstlichen Materialklassen)

	C4.5		NBayes		kNNearest		LogitBoost		SVM (RBF)		SVM (linear)		RF		ELM		svmPoly		MLP	
	EER [%]	GER [%]	EER [%]	GER [%]	EER [%]	GER [%]	EER [%]	GER [%]	EER [%]	GER [%]	EER [%]	GER [%]	EER [%]	GER [%]	EER [%]	GER [%]	EER [%]	GER [%]	EER [%]	GER [%]
Beton	76,6		66,4		79,3		85,0		92,6		93,7		90,7		84,0		93,9		93,3	
Gips	79,1		72,4		82,6		84,3		88,5		93,4		87,5		86,8		93,3		88,2	
Granit	74,0		46,4		83,8		94,3		95,1		97,0		88,9		67,7		97,8		93,1	
Kalksandstein	82,6	82,5	76,1	71,5	75,3	77,1	84,2	87,6	93,8	92,8	95,5	95,4	90,8	90,6	96,2	82,3	95,1	94,7	93,9	92,9
Leichtbeton	86,8		71,8		80,9		89,9		95,1		97,5		91,2		73,9		96,9		95,3	
Porenbeton	91,6		82,0		84,2		92,7		98,1		98,5		97,9		98,6		98,7		98,2	
Ziegel dicht	89,8		77,2		73,6		92,2		93,5		97,2		85,2		74,0		89,5		87,7	
Ziegel porös	73,7		63,3		60,4		83,0		83,7		90,5		84,2		69,3		89,4		85,4	

Für andere Klassifikatoren wurden die jeweiligen Parameteroptimierungen auch durchgeführt. Die Klassifikationsleistungen sind in der Tabelle 14 dargestellt.

Tabelle 14: Erkennungsrate (Untersuchung mit 5 Oberklassen)

	SVM (linear)		SVM (RBF)		LogitBoost		RF		svmPoly		MLP	
	EER [%]	GER [%]	EER [%]	GER [%]	EER [%]	GER [%]	EER [%]	GER [%]	EER [%]	GER [%]	EER [%]	GER [%]
Klasse 1	89,4		91,9		82,2		91,6		92,7		94,9	
Klasse 2	97,4		98,2		95,0		97,7		97,1		97,2	
Klasse 3	94,0	93,9	94,3	95,1	86,8	88,7	92,3	94,4	94,9	96,1	95,7	96,5
Klasse 4	95,7		96,6		91,4		96,4		98,1		97,7	
Klasse 5	90,5		92,2		83,7		90,0		94,5		94,1	

Die gewählten Klassifikatoren zeigen relativ hohe Erkennungsraten über 90%. Nur *LogitBoost* zeigt eine geringere Erkennungsrate um 88,7%. *Random Forest* zeigt gute Ergebnisse mit einer GER um 94%, was kleiner als bei svmPoly und MLP ist. Der Klassifikator SVM mit linearem Kern zeigt eine schlechtere Generalisierung im Vergleich zum ersten Test mit 51 Subklassen, was als Überanpassung im Fall von wenigen Testobjekten beschrieben werden kann (die Anzahl der Objekte pro Klasse im Fall von 51 Subklassen ist geringer als im Fall mit den 5 bzw. 8 Klassen). Die Klassifikatoren svmPoly und MLP zeigen eine gute Generalisierungsfähigkeit und erreichen eine GER um 96,1% und 96,5% entsprechend. Die meisten Klassifikatoren haben Schwierigkeiten bei der Erkennung der Klasse 1 mit einer Erkennungsrate unter 95%.

Die Klassifikationsleistungen auf dem Datensatz mit 8 Materialklassen sind in Tabelle 15 dargestellt.

Tabelle 15: Erkennungsrate (Untersuchung mit 8 Materialklassen)

	SVM (linear)		SVM (RBF)		LogitBoost		RF		svmPoly		MLP	
	EER [%]	GER [%]	EER [%]	GER [%]	EER [%]	GER [%]	EER [%]	GER [%]	EER [%]	GER [%]	EER [%]	GER [%]
Beton	89,3		90,5		79,8		88,4		92,8		94,6	
Gips	92,1		92,0		83,2		90,1		93,9		93,6	
Granit	97,5		98,0		95,0		94,7		96,7		96,6	
Kalksandstein	94,0	94,8	93,6	94,8	85,3	88,7	91,2	93,5	95,0	96,3	95,7	96,6
Leichtbeton	93,6		95,9		86,0		94,6		96,6		96,7	
Porenbeton	98,8		96,3		92,6		96,5		99,0		98,6	
Ziegel dicht	97,7		98,0		96,6		97,1		97,6		98,0	
Ziegel porös	99,4		98,0		98,4		97,5		99,2		98,8	

Laut dieser Ergebnisse setzt sich der Trend vom letzten untersuchten Datensatz fort. Die schlechteste Erkennungsrate um 88,7% zeigt der Klassifikator LogitBoost. Der Klassifikator

Random Forest zeigt eine geringere Erkennungsrate im Vergleich zu SVM und MLP. Eine Schwachstelle bei diesem Ensemble an Entscheidungsbäumen ist die Klasse Beton mit einer Erkennungsrate von 88,4%, bei anderen Klassen sind bessere Leistungen zu beobachten. Die Klassifikatoren auf der Basis der SVM zeigen einen kleinen Anstieg der Erkennungsrate. Die svmPoly und MLP erreichen die beste Leistung auf diesem Datensatz mit der Gesamterkennungsrate von über 96% und Einzelerkennungsraten von über 95% für alle Materialklassen außer Beton (92,8% und 94,6% entsprechend) und Gips (93,9% und 93,6% entsprechend).

Unter Berücksichtigung der Ergebnisse kann angenommen werden, dass die Klassifikatoren svmPoly und MLP die beste Leistung auf dem Bilddatensatz zeigen und werden daher für die weiteren Untersuchungen benutzt.

12.4. Untersuchung des Einflusses des Merkmalsselektionsverfahrens auf die Klassifikationsperformance

Eine korrekte Auswahl der Merkmale ist sehr wichtig für die Lösung der Klassifikationsaufgabe. Die Daten können informationsirrelevantes Rauschen in Form einer großen Anzahl irrelevanter Merkmale enthalten, Merkmale mit geringer Varianz oder mit starken Korrelationen untereinander. Hier kommen Merkmalsselektionsverfahren zum Einsatz, welche den Informationsgehalt, die Varianz, die Korrelationen und andere Bewertungskriterien für die nachfolgende Auswahl des optimalen Merkmalssatzes analysieren.

Es existieren zahlreiche Methoden für die Auswahl der besten Merkmale. Man unterscheidet zwischen Filter-, Wrapper- und Embedded-Merkmalsselektionsverfahren. Die Filter-Verfahren sind unabhängig vom Klassifikator und haben einen niedrigen Rechenaufwand, was zu einer leichten Implementierung in den Klassifikationsprozess führt. Diese Verfahren wurden zuerst getestet.

Unter Einsatz der Klassifikatoren svmPoly, MLP, Random Forest und LogitBoost (die ersten zwei wurden wegen der besten Leistungen auf dem Datensatz gewählt, die letzten zwei stellen andere Klassifikationsalgorithmen zum Vergleich dar) aus der Caret-Bibliothek wurden die Filterverfahren *ReliefF*, *chiSquare-* und *InfoGain-Ranking* geprüft. Für die Untersuchungen mit Merkmalsselektionsverfahren wurden 60% des Datensatzes als Trainingsdatensatz und als Basis für die Selektion verwendet und 40% als Testdatensatz. Diese Trennung wurde zehnmal zufällig wiederholt. Die Ergebnisse wurden dann gemittelt. Die Ergebnisse sind in Abb. 26, 29, 32 für den Datensatz mit 5 Oberklassen und in Abb. 27, 30 und 33 für den Datensatz mit 8 Materialklassen dargestellt.

Die Nutzung der geringeren Anzahl an Merkmalen beeinflusst unter Anwendung von dem InfoGain- und Chi-Square-Filter die Gesamterkennungsrate stark. Der Bereich von 15 bis 35 Merkmalen (im Fall des svmPoly-Klassifikators bis 75 Merkmalen) zeigt, dass der Klassifikator nicht genügend Informationen hat, um eine präzise Vorhersage zu machen, was zu schlechteren Erkennungsraten mit 87% für MLP, 84% für Random Forest, 80% für LogitBoost und 77% für svmPoly bei der Nutzung von 15 durch InfoGain-Filter ausgewählten Merkmale führt. Die Grafik zeigt eine Stabilisierung ab 75 Merkmalen auf eine GER über 96% für MLP und svmPoly, über 93% für Random Forest und über 88% für LogitBoost. Im Bereich von 95 bis 135 Merkmalen befindet sich ein lokales Maximum und eine weitere Erhöhung der Merkmalsanzahl leistet keine weitere Verbesserung.

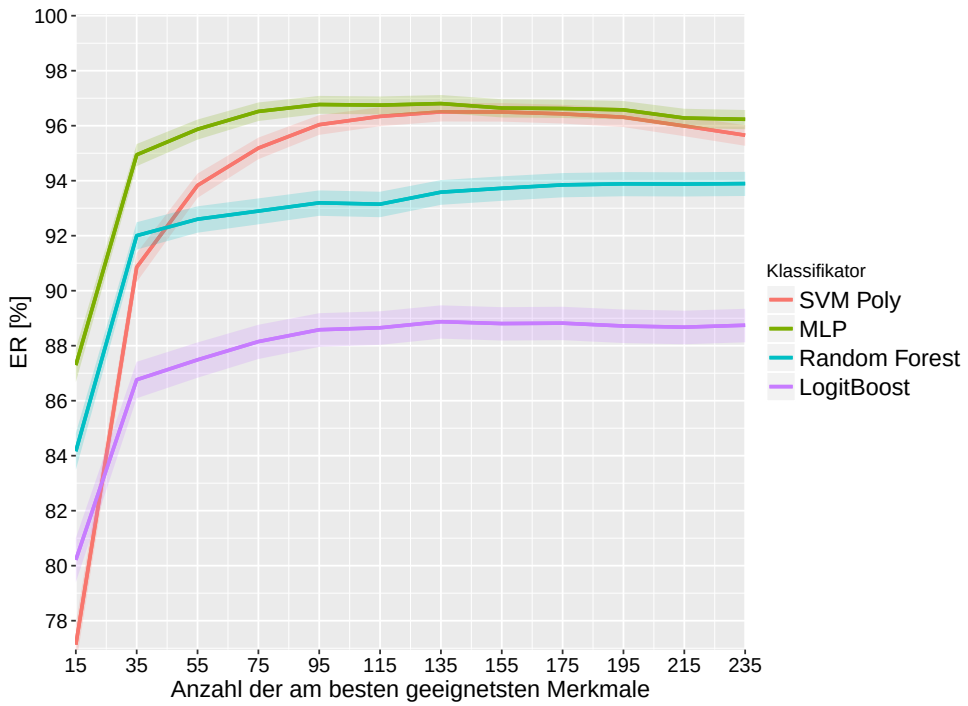

Abbildung 26: Leistung der Klassifikatoren auf den besten, mit *InfoGain-Ranking* gewählten Merkmale (5 Oberklassen)

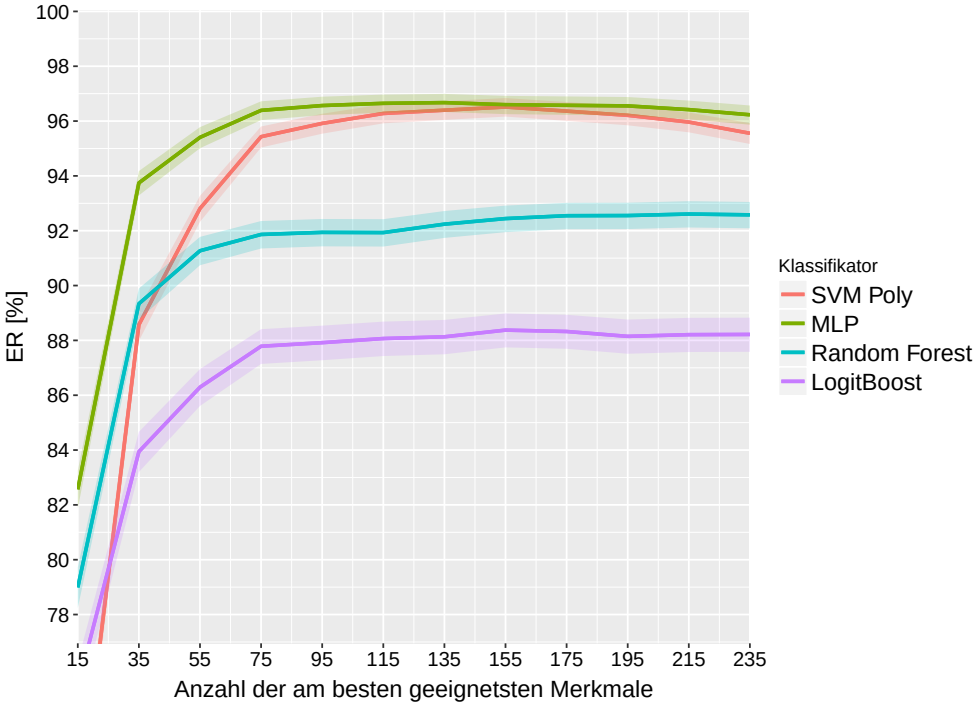

Abbildung 27: Leistung der Klassifikatoren auf den besten, mit *InfoGain-Ranking* gewählten Merkmale (8 Materialklassen)

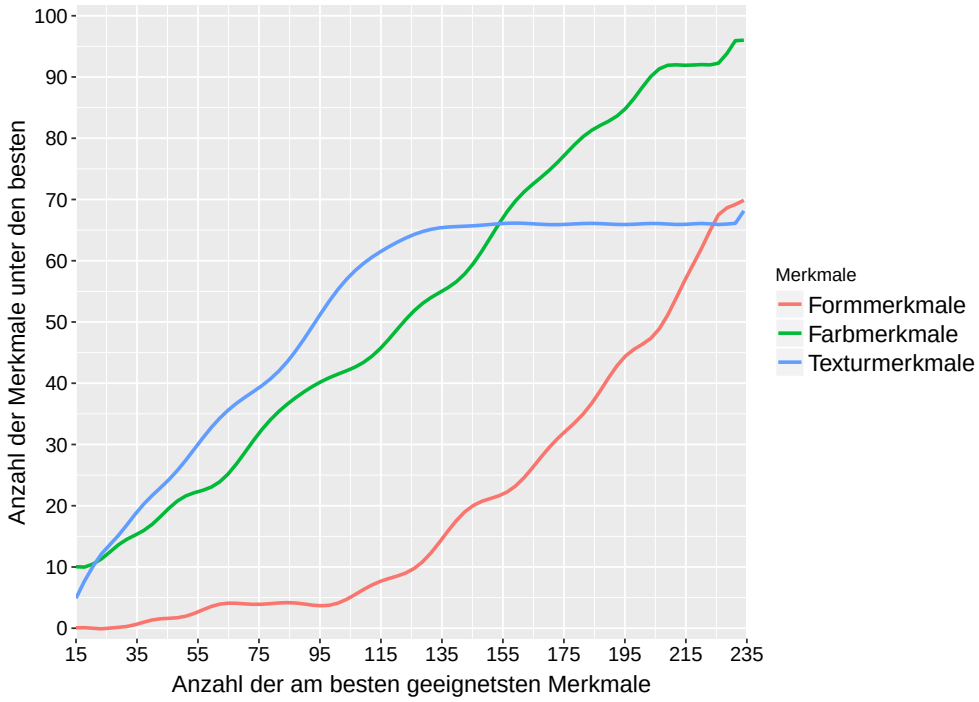

Abbildung 28: Anzahl der bestimmten Merkmale unter den besten, mit *InfoGain-Ranking* gewählten Merkmale (8 Materialklassen)

In der Abbildung 28 sind die Anzahl der bestimmten Merkmale unter den besten dargestellt. Aus der Graphik ist es ersichtlich, dass die Textur- und Farbmerkmale bis 115 Merkmale eine Mehrzahl haben (62 Textur- und 46 Farbmerkmale unter den 115 besten laut InfoGain-Filter). Die Erkennungsleistungen der Klassifikatoren haben ein Plateau ab 115 Merkmale. Das bedeutet, dass diese ersten 115 Merkmale mit höchster Signifikanz eine entscheidende Rolle für Erkennungsleistungen haben.

Wie bei der Anwendung von InfoGain-Filter haben Textur- und Farbmerkmale den Haupteinfluss auf die Klassifikationsleistungen laut chiSquare-Ranking. Der Anteil der Merkmale unter den besten ist genauso hoch wie beim InfoGain (Abb. 31).

Der Unterschied zwischen InfoGain- und chiSquare-Filter besteht darin, dass alle Klassifikatoren unter Anwendung der 15 besten, mit chiSquare-Ranking ausgewählten Merkmale bessere Leistungen (um 2% höhere Gesamterkennungsrate) zeigen. Trotzdem weisen die Klassifikatoren bei der geringeren Merkmalsanzahl die Unteranpassung auf.

Gemäß der Ergebnisse der Klassifikation unterscheidet sich die optimale Merkmalsanzahl von Klassifikator zu Klassifikator. Bei etwa 95 Merkmalen unter Anwendung des *InfoGain-Ranking* bzw. *chiSquare-Ranking* erreicht der Klassifikator MLP eine GER von 96,8% . Die optimale Merkmalsanzahl liegt bei 135 für svmPoly (GER von 96,5%) und für LogitBoost (GER von 89%), bei 195 für Random Forest (GER von 94%). Es ist wert zu erwähnen, dass die Anwendung des InfoGain-Filters im Zusammenhang mit Random Forest fast sinnlos ist, weil der Klassifikator eine Embedded-Merkmalsselektion nutzt, welche auch auf der Berechnung der Entropie von Merkmalen basiert.

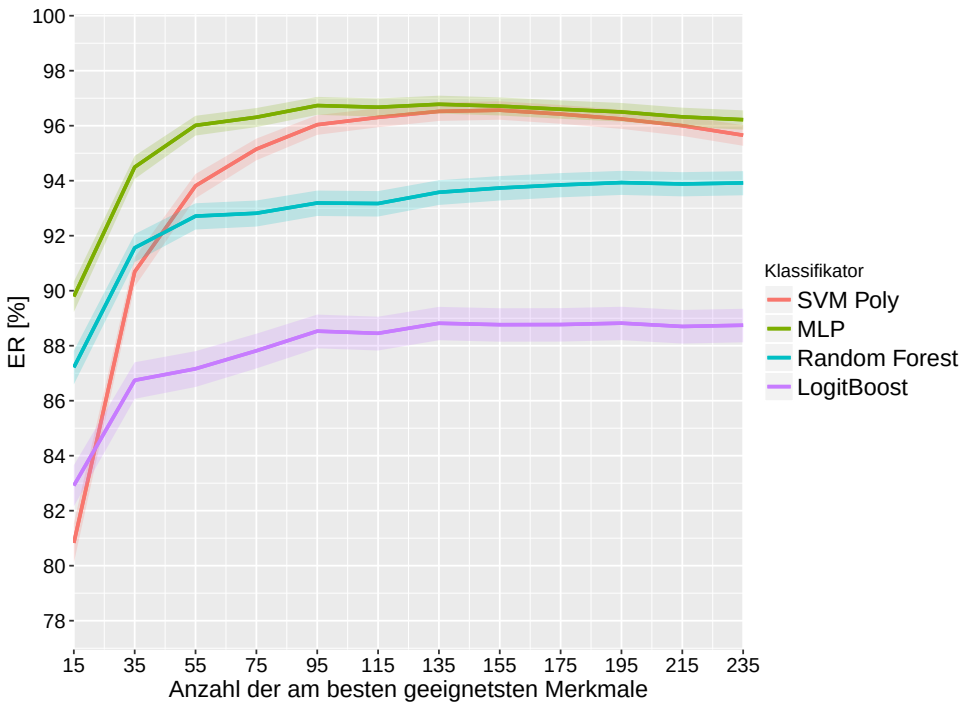

Abbildung 29: Leistung der Klassifikatoren auf den besten, mit *chiSquare-Ranking* gewählten Merkmalen (5 Oberklassen)

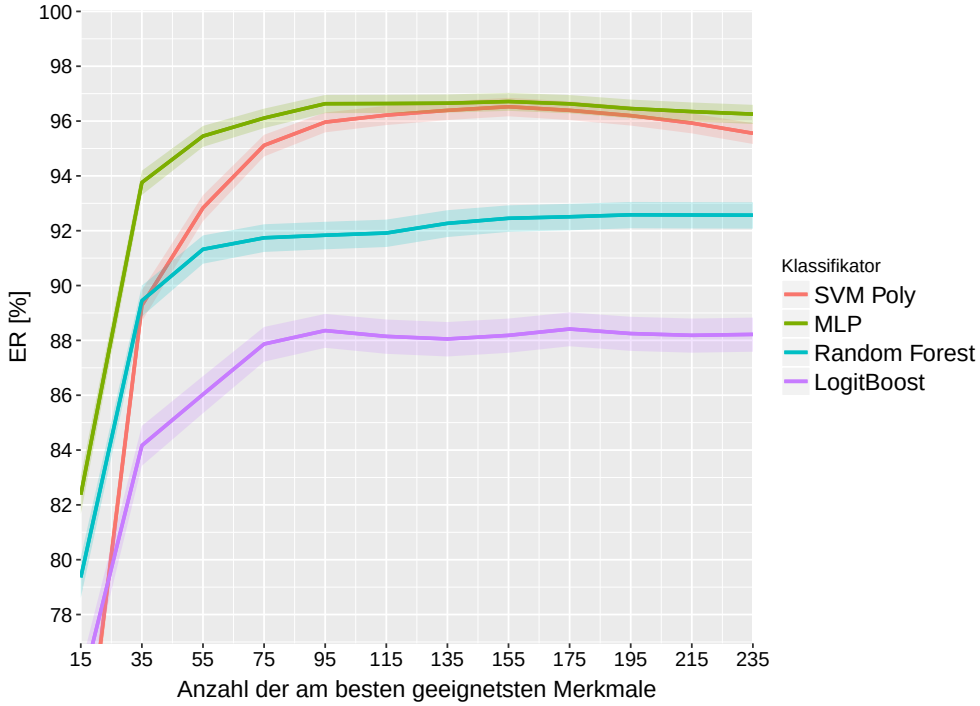

Abbildung 30: Leistung der Klassifikatoren auf den besten, mit *chiSquare-Ranking* gewählten Merkmalen (8 Materialklassen)

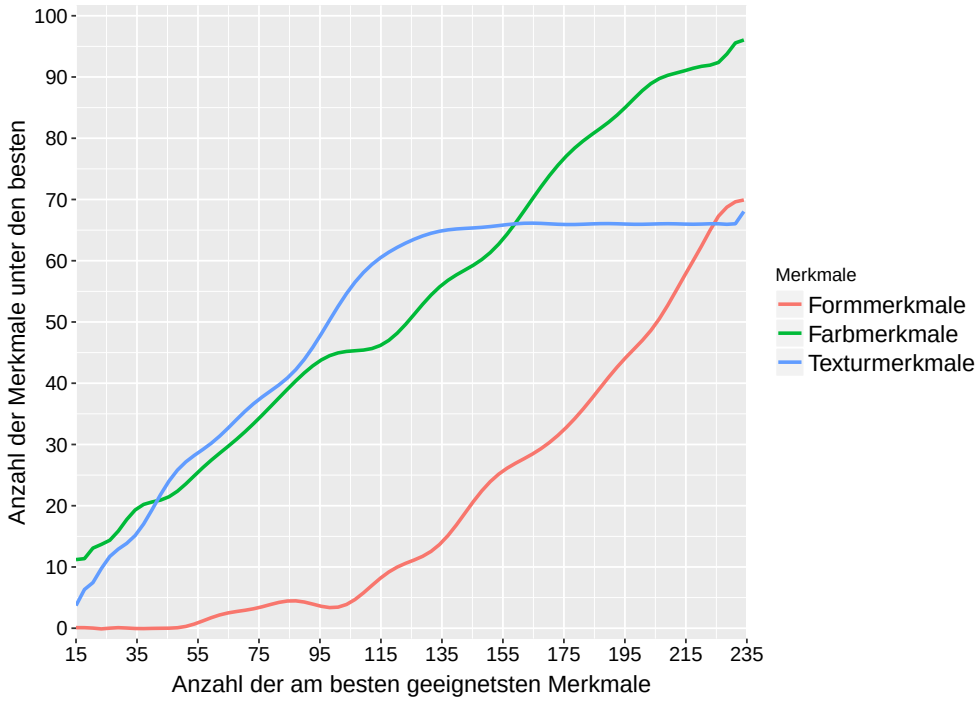

Abbildung 31: Anzahl der bestimmten Merkmale unter den besten, mit *chiSquare-Ranking* gewählten Merkmale (8 Materialklassen)

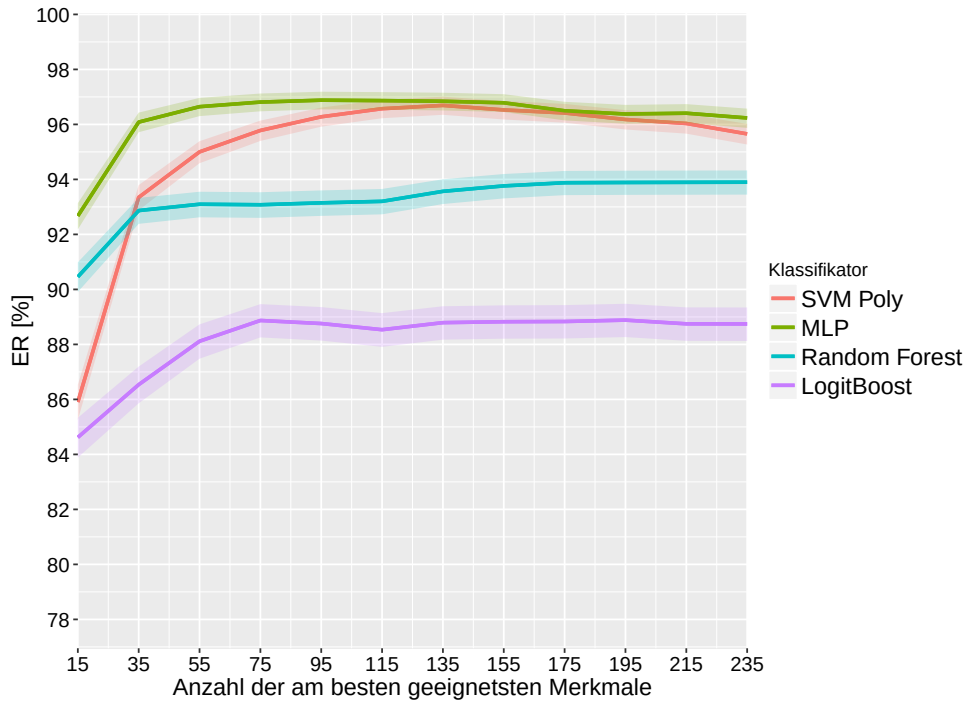

Abbildung 32: Leistung der Klassifikatoren auf den besten, mit *ReliefF-Filter* gewählte Merkmalen (5 Oberklassen)

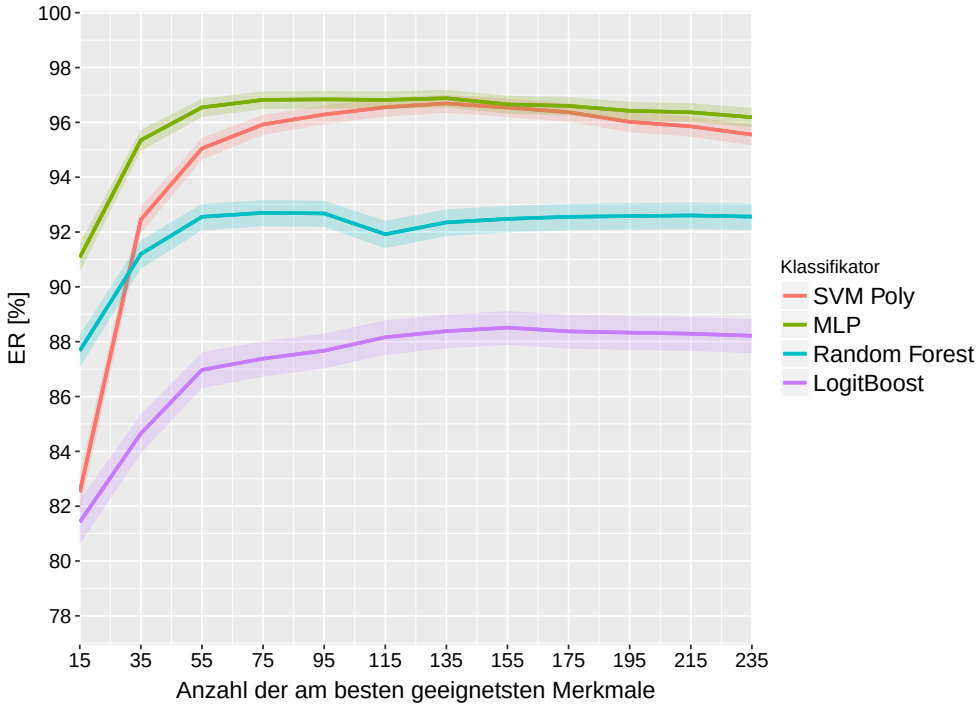

Abbildung 33: Leistung der Klassifikatoren auf den besten, mit *ReliefF-Filter* gewählten Merkmalen (8 Materialklassen)

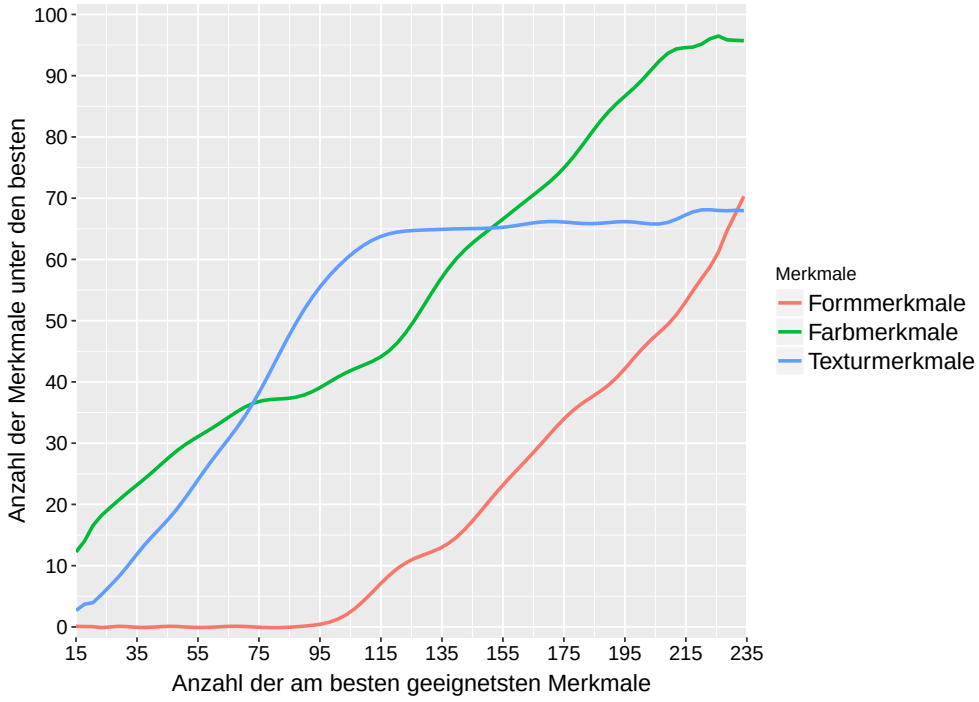

Abbildung 34: Anzahl der bestimmten Merkmale unter den besten, mit *ReliefF-Filter* gewählten Merkmale (8 Materialklassen)

Die Anwendung des ReliefF-Filters zeigt etwas bessere Ergebnisse. Der Unterschied liegt

darin, dass weniger Merkmale notwendig sind, um ein Plateau zu erreichen. Schon bei der Anwendung von 35 Merkmalen zeigt MLP die Gesamterkennungsrate höher als 96% und Random Forest zeigt eine GER um 92,4%, was nur um 1% weniger als die höchste erreichte Gesamterkennungsrate (93,4%) ist. Die optimale Anzahl der Merkmale laut *ReliefF*-Filter liegt bei 75-95 für MLP, was eine GER von 96,8% ergibt. Es ist möglich, auch die Anzahl weiterer als bei *Chi-Square* oder *Info-Gain*-Merkmalsselektion ohne großen Verlust an Genauigkeit zu reduzieren. Die 55 besten Merkmale laut ReliefF-Filter in der Kombination mit dem MLP-Klassifikator führen zu einer GER um 96,3%.

Der ReliefF-Filter bevorzugt am Anfang im Gegensatz zum InfoGain-Filter und chiSquare-Ranking die Farbmerkmale vor den Texturmerkmalen (siehe Abbildung 34). Die Verteilung der Merkmale unter den 115 besten bleibt unverändert - die Texturmerkmale haben den größten Anteil und sind nah an den Farbmerkmalen.

Die Klassifikatoren svmPoly und MLP weisen sowohl die Unteranpassung bei der geringeren Merkmalsanzahl als auch die Überanpassung bei der großen Merkmalsanzahl auf. Die optimale Merkmalsanzahl liegt für der Klassifikator svmPoly bei 135 Merkmalen und bei 95 für MLP ohne Abhängigkeit von der Art der Filterverfahren. Eine weitere Erhöhung der Merkmalsanzahl führt zur Verschlechterung der Leistungen. Die Klassifikatoren Random Forest und LogitBoost weisen nur die Unteranpassung bei geringerer Merkmalsanzahl auf und erreichen die besten Ergebnisse bei der Anwendung des ganzen Merkmalssatzes.

Wegen des fehlenden Zusammenhangs zum verwendeten Klassifikator kann der gewählte Merkmalssatz nicht optimal sondern lediglich suboptimal sein. Ein weiterer Nachteil besteht darin, dass ein univariater Filter bei der Bewertung nur ein Merkmal unabhängig von anderen berücksichtigt, was zur Auswahl hoch korrelierter Merkmale führen kann. Dieser Nachteil ist bei multivariaten Filterverfahren nicht vorhanden. Dieser Filtertyp berücksichtigt den gemeinsamen Einfluss der Merkmale auf einander und auf die Endergebnisse.

Die *Correlation feature selection* (CFS) ist ein Verfahren für die Suche der für die Klassifikation wichtigen Merkmale, die nicht miteinander korreliert sind.
Folgende 18 Merkmale wurden aus dem Datensatz gewählt:

FormfaktorKompaktheit	*HomogenityS*
EnergyI	*HomogenityI*
ContrastI	*AnisotropyH*
GammaS	*GammaH*
GammaI	*ZetaS*
FuzzyEntropyS	*ZetaH*
FuzzyEntropyH	*MaxS*
MinI	*MaxI*
DeviationI_ee	*MeanH_rl*

Die Anwendung des Klassifikators MLP auf dem reduzierten Datensatz ergibt eine Erkennungsrate für Klasse 1 von 91,4%, Klasse 2 von 96,8%, Klasse 3 von 91,3%, Klasse 4 von 97,3%, Klasse 5 von 85,6% und eine GER von 94,3%.

Vergleich der Klassifikationsperformance unter Anwendung von Merkmalsselektionsverfahren

Mit jedem Merkmalsselektionsverfahren hat der Klassifikator MLP die besten Leistungen im Vergleich zu den anderen Klassifikatoren mit diesen Verfahren auf den Datensätze mit 5 Oberklassen und 8 Materialklassen gezeigt. Die Ergebnisse sind in den Tabellen 16 und 17 entsprechend dargestellt.

Tabelle 16: Erkennungsrate (Untersuchung mit 5 Oberklassen unter Anwendung von Merkmalsselektionsalgorithmen)

	MLP (InfoGain 135 Merkmale)		MLP (ChiSquare 135 Merkmale)		MLP (ReliefF 95 Merkmale)		MLP (CFS 18 Merkmale)	
	EER [%]	GER [%]	EER [%]	GER [%]	EER [%]	GER [%]	EER [%]	GER [%]
Klasse 1	96,8		96,8		95,3		91,5	
Klasse 2	99,0		98,9		98,3		96,8	
Klasse 3	97,9	96,8	97,8	96,8	95,8	96,8	91,3	94,3
Klasse 4	97,9		97,8		98,0		97,3	
Klasse 5	97,0		97,2		94,4		85,6	

Tabelle 17: Erkennungsrate (Untersuchung mit 8 Materialklassen unter Anwendung von Merkmalsselektionsalgorithmen)

	MLP (InfoGain 135 Merkmale)		MLP (ChiSquare 155 Merkmale)		MLP (ReliefF 135 Merkmale)		MLP (CFS 30 Merkmale)	
	EER [%]	GER [%]	EER [%]	GER [%]	EER [%]	GER [%]	EER [%]	GER [%]
Beton	94,7		94,8		95,0		92,4	
Gips	93,9		94,1		94,2		90,7	
Granit	97,5		97,2		96,6		95,5	
Kalksandstein	95,1	96,7	95,2	96,7	95,3	96,9	93,1	95,0
Leichtbeton	96,7		96,8		97,1		95,0	
Porenbeton	98,9		98,7		99,0		98,0	
Ziegel dicht	98,0		98,4		98,4		97,3	
Ziegel porös	99,0		99,0		99,0		98,3	

Alle drei Filterverfahren (InfoGain, chiSquare, ReliefF) haben ähnliche Ergebnisse auf den beiden Datensätze gezeigt. Die Anwendung von ReliefF lässt eine etwas kleinere Anzahl von Merkmalen im Vergleich zu zwei anderen Filterverfahren auf dem Datensatz mit 5 Oberklassen nutzen, aber das ergibt die höhere Erkennungsrate nur für Klasse 4 - die anderen einzelnen Erkennungsraten haben sich verkleinert.

Obwohl der Correlation feature selection Algorithmus den kleinsten Merkmalssatz zwischen allen getesteten Algorithmen ergibt, sind die Ergebnisse auf den beiden Datensätzen schlechter im Vergleich zu anderen angewendeten Verfahren.

Das InfoGain-Merkmalsselektionsverfahren hat den geringsten Rechenaufwand zusammen mit dem chiSquare-Verfahren und hat eine der besten Leistungen auf den beiden Datensätzen gezeigt, weswegen die Anwendung des Verfahrens für die Klassifikationsaufgabe optimal ist.

12.5. Hauptkomponentenanalyse von Bildinformationen

Die Hauptkomponentenanalyse ist ein Verfahren für die Dimensionsreduktion/Merkmalsextraktion und kann für diese Aufgabe die Anzahl der Merkmale verringern bzw. neue Merkmale extrahieren, während die relevante Information bewahrt wird.

Die Hauptkomponentenanalyse wurde auf den Bilddatensätzen mit 5 Oberklassen und 8 Materialklassen nach einer vorherigen Zentrierung und Normierung angewandt. Die berechnete Hauptkomponenten wurden dann für das Training und Test der Klassifikatoren verwendet. Die Leistungen der Klassifikatoren sind in der Abbildungen 35 und 36 dargestellt.

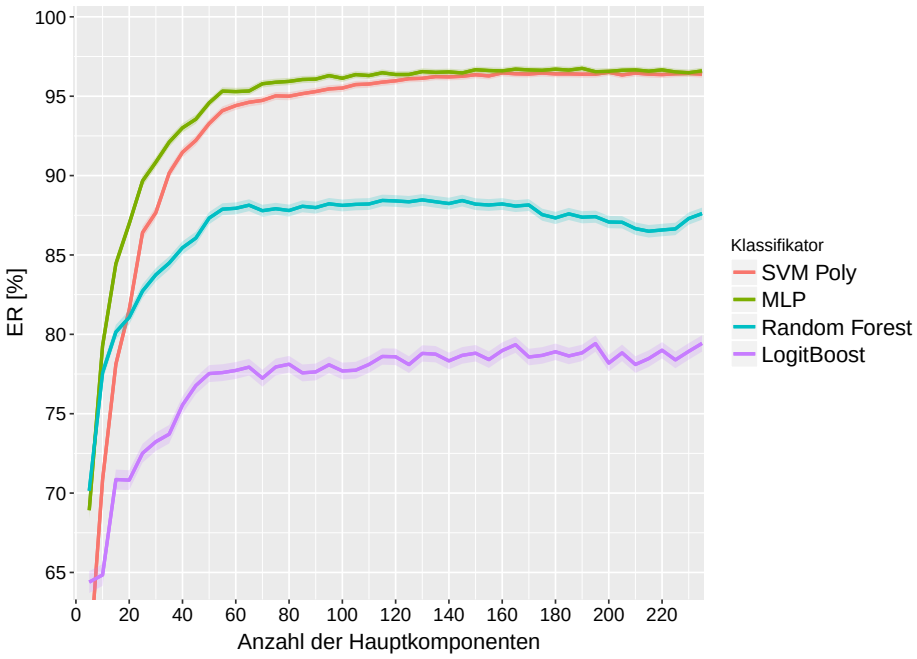

Abbildung 35: Leistung der Klassifikatoren auf den Hauptkomponenten (5 Oberklassen)

Die Anwendung von weniger als 40 Hauptkomponenten weist schlechte Erkennungsraten (unter 90% für svmPoly und MLP, unter 82% für Random Forest und LogitBoost) auf. Die Klassifikatoren erreichen ein Plateau ab 70 Hauptkomponenten für beide Datensätze. Die Klassifikatoren svmPoly und MLP zeigen dabei die Ergebnisse, die den besten erreichten Leistungen unter Anwendung von Filter-Verfahren ähnlich sind (über 96%). Die Klassifikatoren Random Forest und LogitBoost zeigen im Gegensatz dazu eine Verschlechterung der Leistungen um 6-8%. Der Grund dafür kann die Anwendung von komplizierteren Merkmalen - Hauptkomponenten sein, die eine Kombination von originalen Bildmerkmalen darstellen und den Einfluss von eingebauten Merkmalsselektionsverfahren, welche in beiden Klassifikatoren eingebunden sind.

Die Ergebnisse der Anwendung der Hauptkomponentenanalyse auf den Bilddatensätze zeigen, dass die Methode zu einer Verschlechterung der Leistungen führt und die Erken-

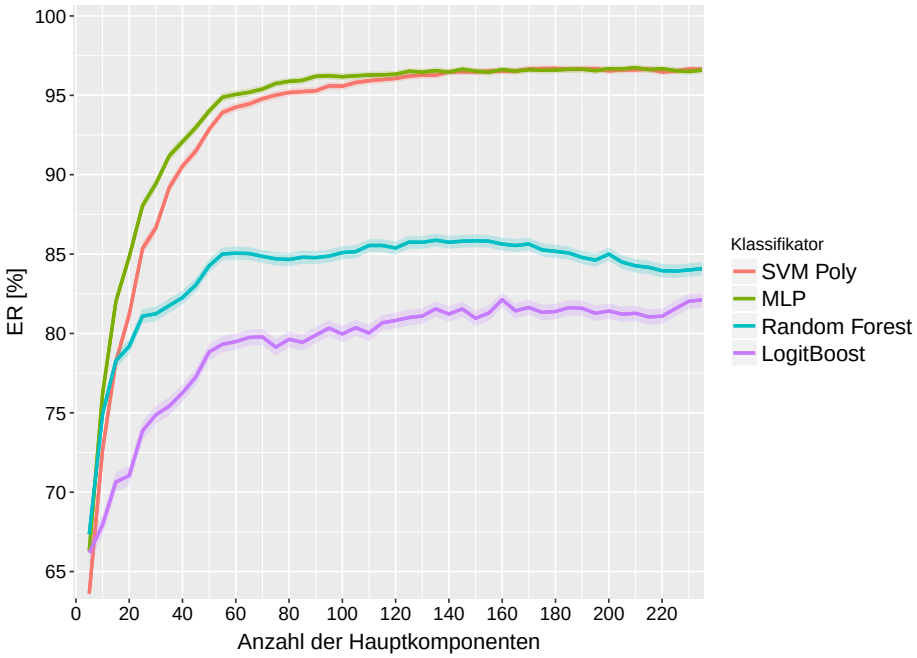

Abbildung 36: Leistung der Klassifikatoren auf den Hauptkomponenten (8 Materialklassen)

nungsaufgabe durch die Einführung von komplizierteren Merkmalen (Hauptkomponenten) erschwert.

13. Anwendung des überwachten maschinellen Lernens auf den Spektraldatensatz

13.1. Analyse der Spektren

Im Spektralbereich von 470-720 nm haben einige Materialien charakteristische Eigenschaften (Abbildung 37), wohingegen andere nur Intensitätsunterschiede zeigen. Charakteristische positive und negative Anstiege haben die Kurven von Ziegel (dicht und porös). Andere Spektralkurven (Beton, Gips, Porenbeton usw.) haben nur Intensitätsunterschiede. Bei der Spektrenaufnahme ist es schwierig, die gleiche Intensität für alle Proben im Rahmen einer Klasse zu bekommen aufgrund der verschiedenen Objektformen und als Folge einer verschiedenen Reflexionsfähigkeit. Chemisch ähnliche Klassen, wie z.B. Leichtbeton und Beton, haben fast gleiche Spektralkurven ohne Intensitätsunterschiede, deswegen reicht der VIS-Bereich für die Erkennung der Subklassen nicht aus.

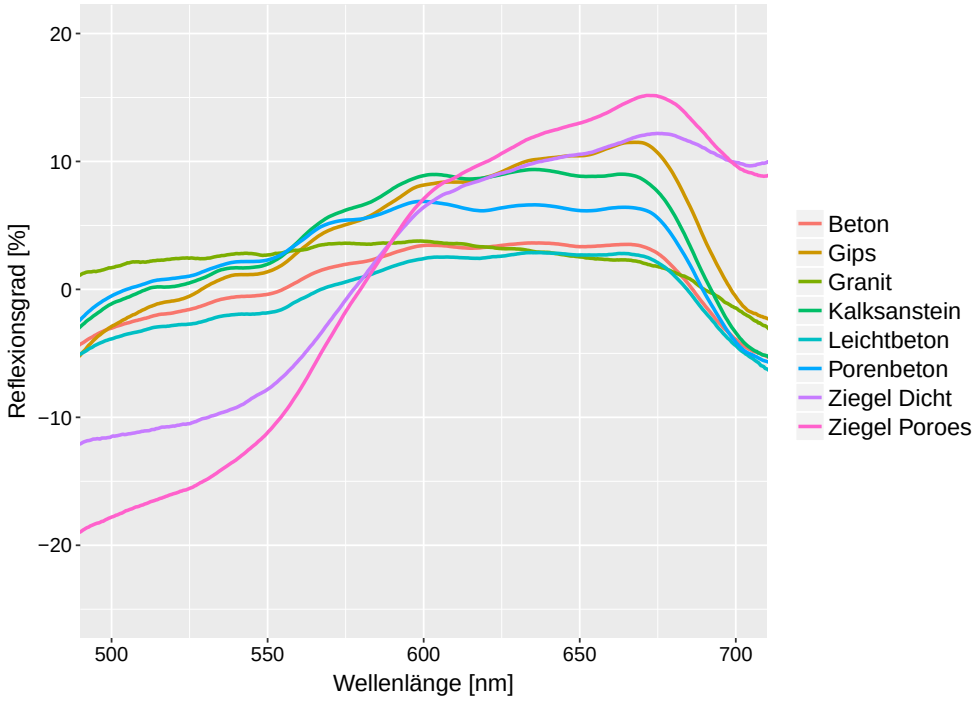

Abbildung 37: Spektren des Bauschutts im VIS-Bereich

In der ersten Ableitung der Spektren ist der Unterschied zwischen einigen Klassen besser erkennbar (siehe Abbildung 38). Dichte und poröse Ziegel haben einen charakteristischen Anstieg ab 550 nm. Im Bereich von 670-720 nm zeichnen sich die Spektren von Gips, Porenbeton und Kalksandstein aus. Kalksandstein scheint trennbar von anderen Klassen zu sein, insbesondere zu den Gips-Spektralkurven, welche einen Abstand bei 650 nm aufweisen. Porenbeton hat eine abweichende Intensität bei 680 nm, die kleiner als bei anderen Klassen ist. Der Unterschied zwischen Gips und Kalksandstein ist deutlich geringer und liegt im Bereich 620-670 nm. Granit zeigt keine charakteristischen Intensitätsänderungen, was auch als ein Merkmal bewertet werden kann. Die Spektralkurven von Beton und Leichtbeton sind fast identisch, so dass die Identifikation sehr schwierig wird.

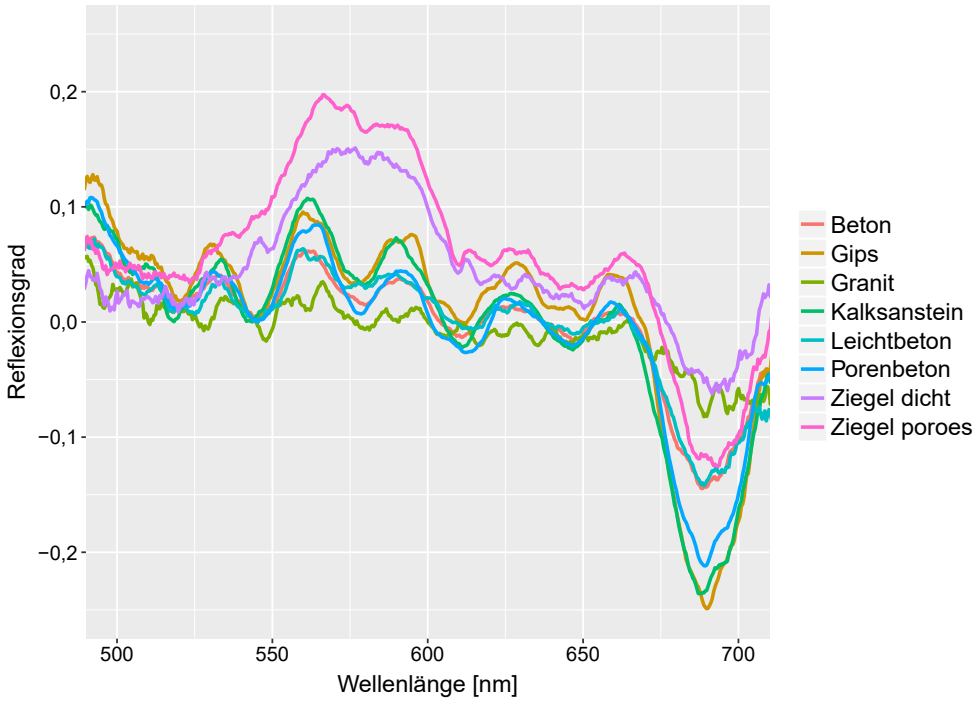

Abbildung 38: Erste Ableitung der Spektren des Bauschutts im VIS-Bereich

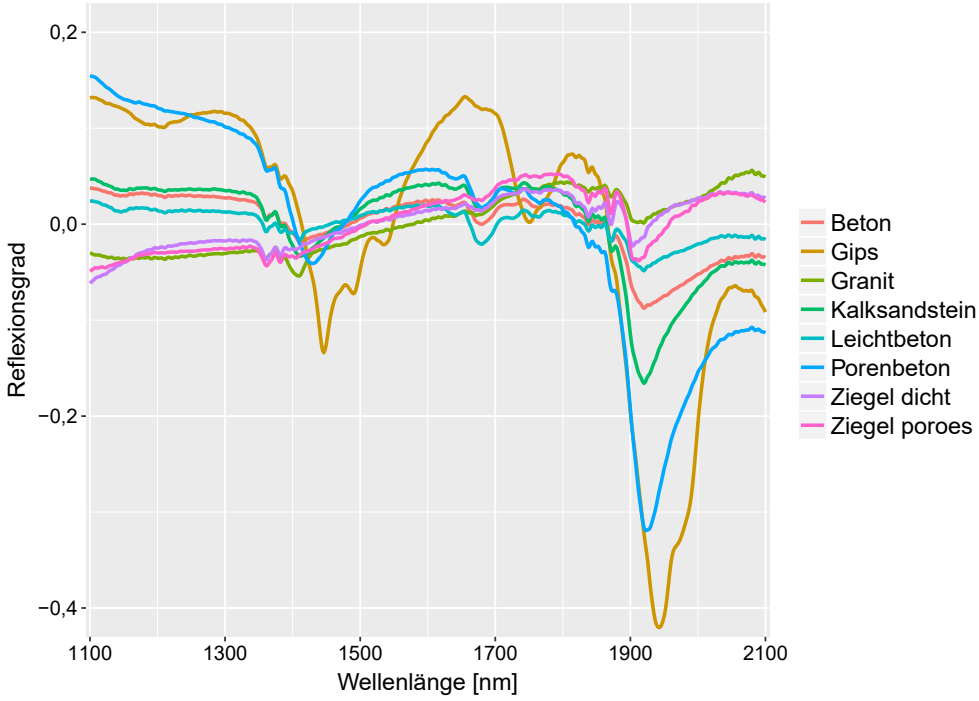

Abbildung 39: Spektren des Bauschutts im IR-Bereich (1100-2100 nm)

Im IR-Bereich zeigen die Klassen deutliche Unterschiede (siehe Abbildung 39). Gips hat starke charakteristische Absorptionsbänder bei 1440, 1750 und 1930 nm und wird gut erkannt. Das Porenbetonspektrum hat starke Absorptionsbänder bei 1430 und 1920 nm, auch ein

schwaches Absorptionsband bei 1680 nm, der Kalksandstein bei 1410, 1680 und 1920 nm. Chemisch ähnliche Klassen, wie Beton und Leichtbeton oder dichte und poröse Ziegel, haben ähnliche Spektralkurven aber der Unterschied liegt in der Intensitätsänderung bei 1920 nm. In der Nähe liegende Spektralkurven, wie Granit, Asphalt, Beton, haben als einzige eine Reihe von Absorptionsbändern, bei denen einige negative Anstiege fehlen und andere dagegen besonders groß sind.

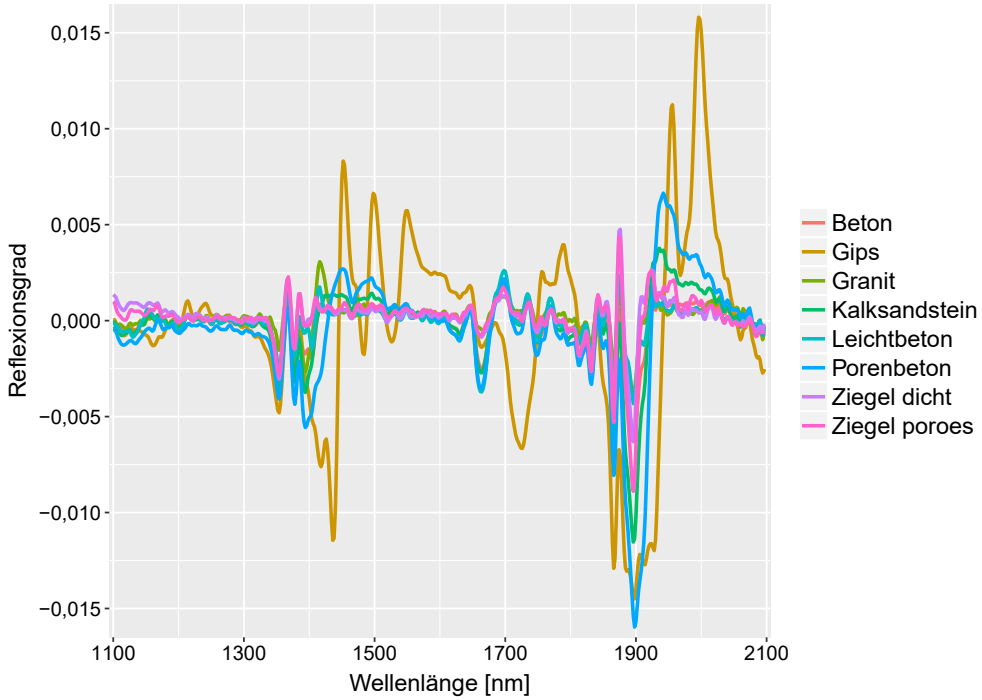

Abbildung 40: Erste Ableitung der Spektren des Bauschutts im IR-Bereich (1100-2100 nm)

Bei der Analyse der ersten Ableitung kann man Gips und Porenbeton sehr gut unterscheiden (siehe Abbildung 40). Gips hat starke Extrema bei 1440, 1460, 1485, 1500, 1550, 1740, 1860, 1900, 1950, 1960 und 2000 nm. Porenbeton zeigt signifikante Eigenschaften bei 1390, 1450, 1490, 1900 und 1940 nm. Kalksandstein zeigt auch charakteristische Eigenschaften, obwohl nicht so stark wie die beiden vorweg genannten Materialien. Granit und Asphalt zeigen wenig Änderungen in der Intensität, was man als ihren spezifischen Fingerabdruck verwenden kann. Beton und Leichtbeton unterscheiden sich in ihrer Intensität bei 1890 nm (siehe Abbildung 41). Porosierter und dichter Ziegel haben unterschiedliche Reflexionsgrade auch bei 1890 nm, obwohl sie weniger Unterschied zueinander haben.

Die oben aufgeführte Bewertung weist nach, dass nur einige Materialien charakteristische Eigenschaften im VIS-Bereich haben, die im Folgenden für die Lösung der Klassifikationsaufgabe angewendet werden können, zu nennen sind hier porosierter und dichter Ziegel, Kalksandstein, Porenbeton und Granit. Die anderen Klassen weisen eine schlechte Erkennbarkeit im VIS-Bereich auf. Im IR-Bereich haben die Baumaterialien mehr Eigenschaften, die für eine Trennung relevant sind. Es ist potentiell möglich alle obengenannten Klassen mittels Spektralanalyse im IR-Bereich zu erkennen.

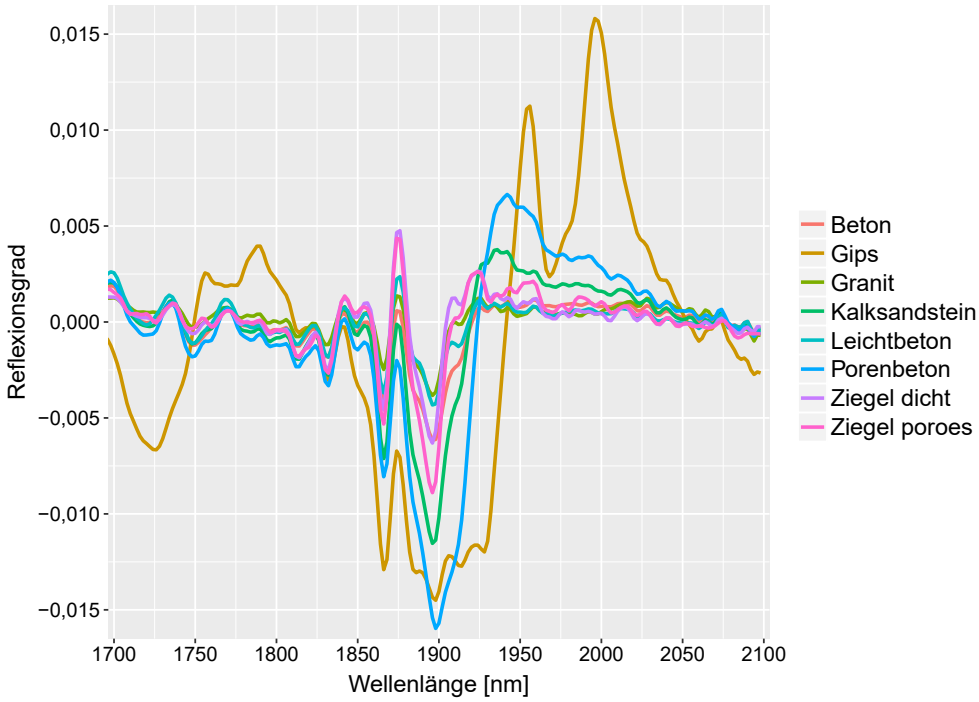

Abbildung 41: Erste Ableitung der Spektren des Bauschutts im IR-Bereich (1700-2050 nm)

13.2. Implementierung Merkmalsextraktion und Merkmalsselektion

Bei der Verwendung der Spektren besteht keine Notwendigkeit für eine externe Merkmalsextraktion, weil das Spektrum selbst schon eine Kombination aus charakteristischen Merkmals-Wellenlängen darstellt, d.h. der Aufnahmeprozess umfasst sowohl die Messung der Spektren als auch die Merkmalsextraktion selbst, obwohl nicht alle aufgenommenen Wellenlängen gleich wichtig für die Lösung der Klassifikationsaufgabe sind. Es gibt jedoch die Möglichkeit, neue Merkmale aus den existierenden Spektren zu extrahieren und die Dimensionalität der Daten damit zu reduzieren. Hierfür wurden sowohl die Hauptkomponentenanalyse als auch die Diskriminanzanalyse angewendet. Es existieren auch andere Möglichkeiten zur Auswahl von Merkmalen mit hoher Bedeutung für die Klassifikation. Dafür wurden Mekmalsselektionverfahren verwendet. Im Gegensatz zur Bildanalyse wurden bei der Spektrenanalyse zuerst die Spektren aufgenommen und dann die Merkmalsselektion durchgeführt.

13.3. Evaluierung geeigneter Klassifikationsverfahren in R

Die Klassifikationsverfahren aus der Bibliothek *caret*, welche bei den Tests auf dem Bilddatensatz ausgewählt wurden, wurden auch auf dem Spektraldatensatz getestet. Es wurden folgende Klassifikatoren verwendet:

- *Random Forest (RF)*
- *Logistic Regression (LogitBoost)*
- *Support Vector Machine mit linear Kernel (svmLinear)*

- *Support Vector Machine mit polynominal Kernel (svmPoly)*
- *Support Vector Machine mit Gaussian (RBF) Kernel (svmRadialSigma)*
- *Multilayer Perceptron (MLP)*

13.3.1. Untersuchung des Einflusses der Datenaufteilung in Trainings- und Testpartitionen

Wegen der kleinen Objektanzahl im Spektraldatensatz kann die Anwendung von Kreuzvalidierungsverfahren zu Überanpassung führen, weil in dem Fall der Testanteil sehr klein und als Folge die Varianz hoch ist. Eine Alternative zur Kreuzvalidierung in dem Fall ist die Anwendung des *Hold-out*-Validierungsschemas. Um ein passendes Trennungsverhältnis zwischen Training- und Testanteil zu finden, wurden die obengenannte Klassifikatoren auf den Spektraldatensätze mit 5 Oberklassen und 8 Materialklassen angewendet. Der Trainingsprozess wird bei den verschiedenen Trainingsanteile von 10 bis 90% von der gesamten Datensatzgröße durchgeführt. Jede Trennung wird zehnmal zufällig wiederholt. Die Ergebnisse sind in der Abbildungen 42 und 43 dargestellt. Zum Vergleich wird dieser Test auch auf dem Bilddatensatz unter Anwendung von *svmPoly*-Klassifikator durchgeführt ("SVM Poly - Bild_DS" in der Legende).

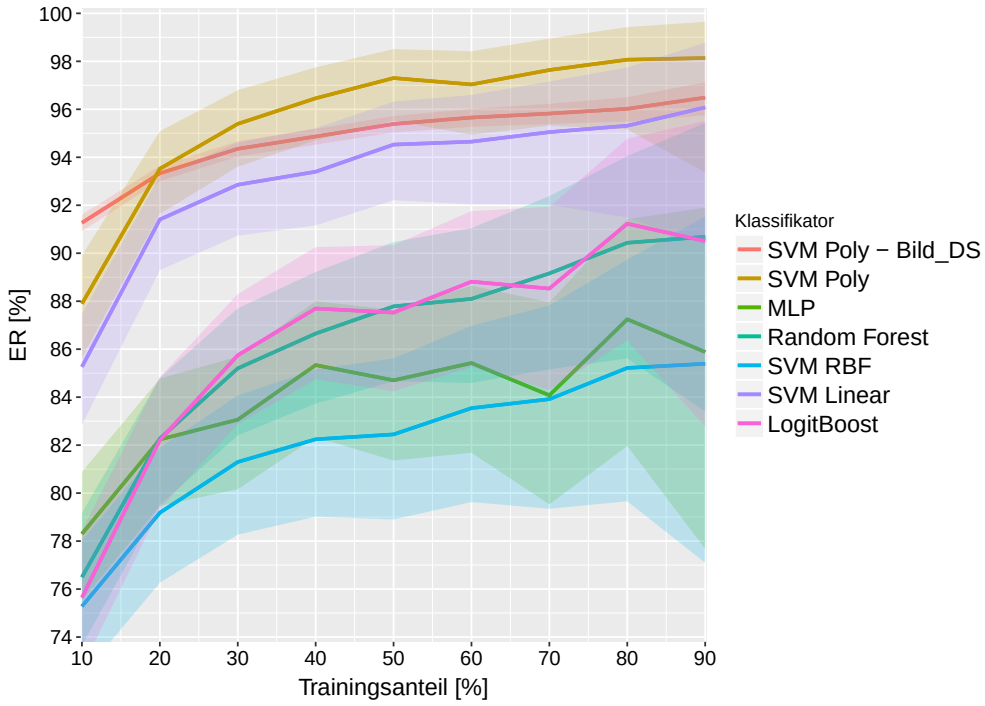

Abbildung 42: Leistungen der verschiedenen Klassifikatoren auf dem Datensatz mit 5 Oberklassen bei unterschiedlichen Trainingsanteilgrößen

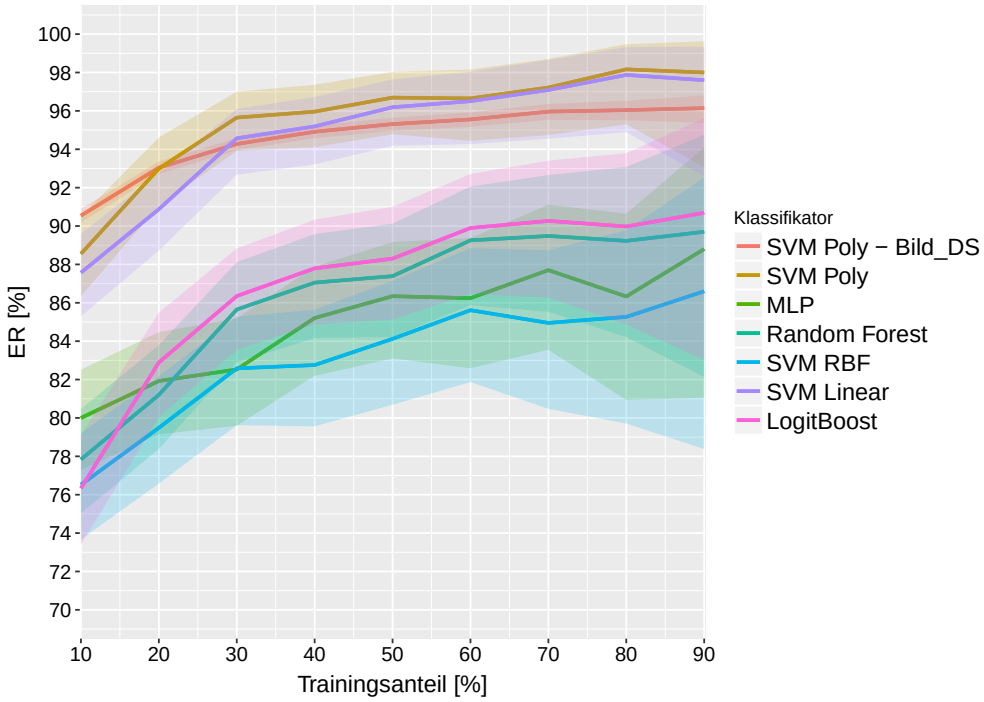

Abbildung 43: Leistungen der verschiedenen Klassifikatoren auf dem Datensatz mit 8 Materialklassen bei unterschiedlichen Trainingsanteilgrößen

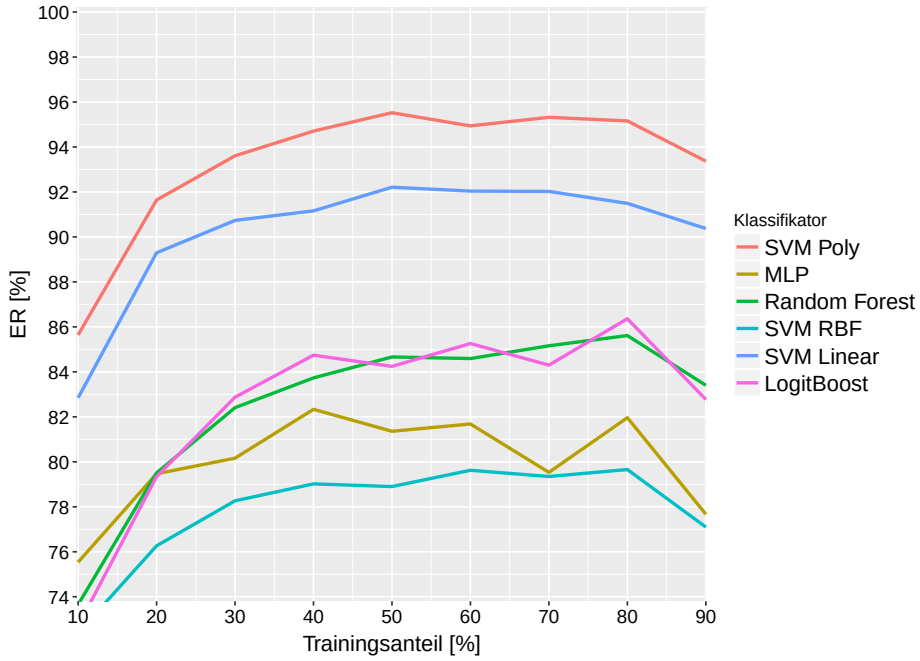

Abbildung 44: Kleinste erreichte Gesamterkennungsraten der verschiedenen Klassifikatoren auf dem Datensatz mit 5 Oberklassen bei unterschiedlichen Trainingsanteilgrößen

Aus den Graphiken ist es ersichtlich, dass die Leistungen auf dem Spektraldatensatz im

Gegensatz zum Bilddatensatz sehr stark von der Trainings-/Testanteilgröße abhängen. Die Abweichung der Erkennungsrate ist auch größer für den Spektraldatensatz, insbesondere bei der Trainingsanteilgrößen über 60%.

Die kleinsten erreichten Erkennungsraten sind einzeln in der Abbildungen 44 und 45 dargestellt. Es ist ersichtlich, dass die Leistungen sich nach dem 50% Anteil für beide Datensätze verschlechtern, was eine Überanpassung bedeuten kann. Anderseits weisen die Klassifikatoren im Bereich unter 30% die Unteranpassung auf. Die Klassifikatoren mit Embedded-Merkmalsselektion wie Random Forest und LogitBoost weisen mehr Robustheit im Vergleich zu Algorithmen ohne integrierte Merkmalsselektion auf, aber zeigen auch ab 60% eine Verschlechterung der Ergebnisse. Der optimale Wert für den Trainingsanteil liegt im Bereich von 30 bis 60% von dem ganzen Datensatz. Für weitere Untersuchungen wurde der Trainingsanteil auf 50% festgelegt.

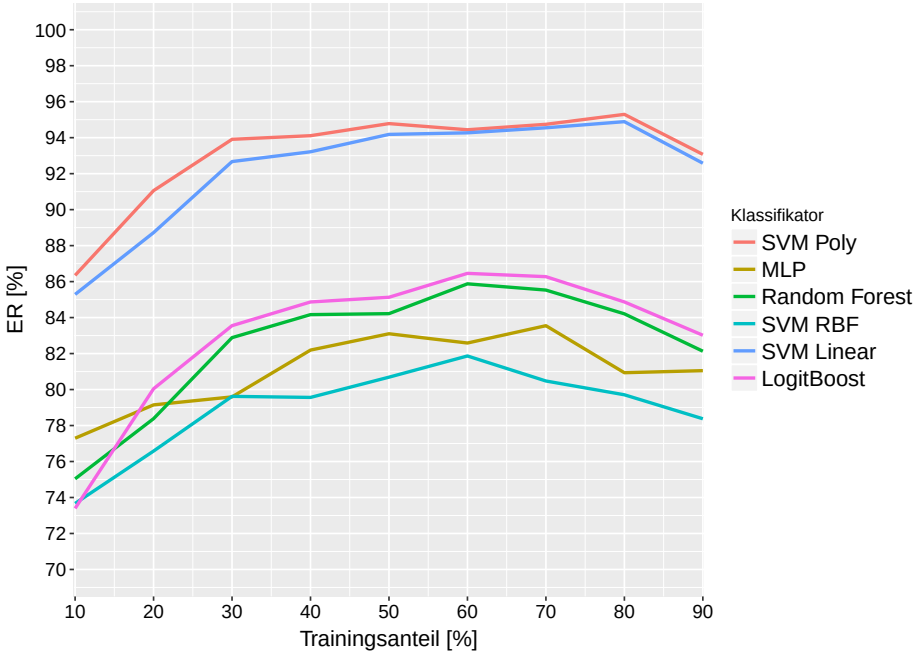

Abbildung 45: Kleinst erreichte Gesamterkennungsraten der verschiedenen Klassifikatoren auf dem Datensatz mit 8 Materialklassen bei unterschiedlichen Trainingsanteilgrößen

13.3.2. Anwendung der Klassifikatoren auf dem Spektraldatensatz

Die Leistungen der Klassifikatoren auf den Datensätze mit 5 Oberklassen (laut DIN-Normen) und mit 8 Materialklassen sind in den Tabellen 18 und 19 dargestellt.

Die Ergebnisse auf dem Datensatz mit 5 Oberklassen zeigen, dass svmPoly die beste Erkennungsrate (97,3%) zwischen alle Klassifikatoren hat. Die Klassifikatoren weisen Schwierigkeiten bei der Erkennung von Klasse 1, Klasse 2 und Klasse 4 auf. Für Klasse 1 und 2 erreichen nur zwei Klassifikatoren eine Erkennungsrate >95% - der Klassifikator svmPoly (95,9% und 97,4%) und svmLinear (95% und 97,4% entsprechend). Mit der Erkennungsrate um 99% für Klasse 4 ist svmPoly der einzige Klassifikator, welcher die Erkennungsrate für diese Klasse über 95% hat, obwohl er etwas schlechtere Ergebnisse für Klasse 5 - 99% im

Vergleich zu 100% bei MLP und SVM mit RBF-Kernel zeigt.

Tabelle 18: Spektrale Untersuchung mit 5 Oberklassen

	LogitBoost nIter = 31		RF mtry = 115		SVM (linear) C = 1		SVM (RBF) sigma=0,15, C = 1		svmPoly degree = 3, scale = 0,1, C = 1		MLP size = 100	
	EER [%]	GER [%]	EER [%]	GER [%]	EER [%]	GER [%]	EER [%]	GER [%]	EER [%]	GER [%]	EER [%]	GER [%]
Klasse 1	82,5		82,8		95,0		86,7		95,9		76,0	
Klasse 2	82,2		80,8		97,4		81,2		97,4		84,3	
Klasse 3	89,0	87,5	89,8	87,8	97,4	94,5	93,1	87,7	94,0	97,3	96,9	84,7
Klasse 4	89,1		89,0		92,6		86,0		99,0		88,5	
Klasse 5	95,5		98,9		99,7		100,0		99,0		100,0	

Alle Klassifikatoren haben ähnliche Leistungen auf den beiden Datensätzen und der Unterschied ist weniger als 1% außer SVM (linear) und MLP. SVM mit linearen Kernel zeigt bessere Leistung auf dem Datensatz mit 8 Materialklassen (96,2%) im Vergleich zu dem Datensatz mit 5 Oberklassen (94,5%). Das MLP zeigt ähnliche Änderungen - 86,3% gegen 84,7%. Die Ergebnisse auf dem Datensatz mit 8 Materialklassen zeigen, welche Materialkomponenten der Oberklassen schwer erkennbar sind. Für Klasse 1, welche die Materialklassen Beton und Granit enthält, stellt die Materialklasse Beton Schwierigkeit dar, weil die Erkennungsraten von Granit für fast alle Klassifikatoren über 90% und die Erkennungsraten von Beton unter 90% (außer svmLinear und svmPoly) liegen. Für Klasse 4 ist die Schwachstelle bei der Materialklasse Leichtbeton - außer svmPoly und svmLinear zeigen alle Klassifikationen einen niedrigere Erkennungsrate als 90%, während die Erkennungsraten für andere Komponenten der Klasse (Porenbeton, Ziegel porös) höher als 90% sind.

Tabelle 19: Spektrale Untersuchung mit 8 Materialklassen

	LogitBoost nIter = 31		RF mtry = 115		SVM (linear) C = 1		SVM (RBF) sigma=0,15, C = 1		svmPoly degree = 3, scale = 0,1, C = 1		MLP size = 100	
	EER [%]	GER [%]	EER [%]	GER [%]	EER [%]	GER [%]	EER [%]	GER [%]	EER [%]	GER [%]	EER [%]	GER [%]
Beton	77,1		77,9		93,3		85,0		94,8		75,6	
Gips	97,0		99,2		100,0		100,0		99,7		99,6	
Granit	94,8		95,3		100,0		100,0		99,2		50,0	
Kalksandstein	88,8	88,3	88,5	87,4	96,8	96,2	92,0	87,2	94,6	96,7	95,3	86,3
Leichtbeton	81,8		82,1		92,5		73,7		96,9		82,2	
Porenbeton	96,9		96,4		99,6		98,2		99,9		99,5	
Ziegel dicht	85,5		76,6		92,7		77,6		94,3		82,2	
Ziegel porös	93,6		90,3		99,2		88,2		97,0		91,9	

Die praktischen Untersuchungen zeigen, dass die Klassifikatoren svmPoly und svmLinear die besten Ergebnisse im Spektraldatensatz haben. Im Gegensatz zum Bilddatensatz zeigt

MLP eine niedrige Erkennungsrate. Andere Klassifikatoren zeigen relativ ähnliche Ergebnisse bei den Erkennungsraten im Bereich von 86% bis 88%. Für weitere Untersuchungen wurden die Klassifikatoren svmPoly und svmLinear (wegen der Leistungen), LogitBoost und Random Forest (zum Vergleich) genutzt.

13.4. Hauptkomponentenanalyse von Spektren

Die Haupkomponentenanalyse (*principal component analysis* - PCA) ist ein Verfahren, welches bei der Suche von Mustern in mehrdimensionalen Daten angewendet werden kann.

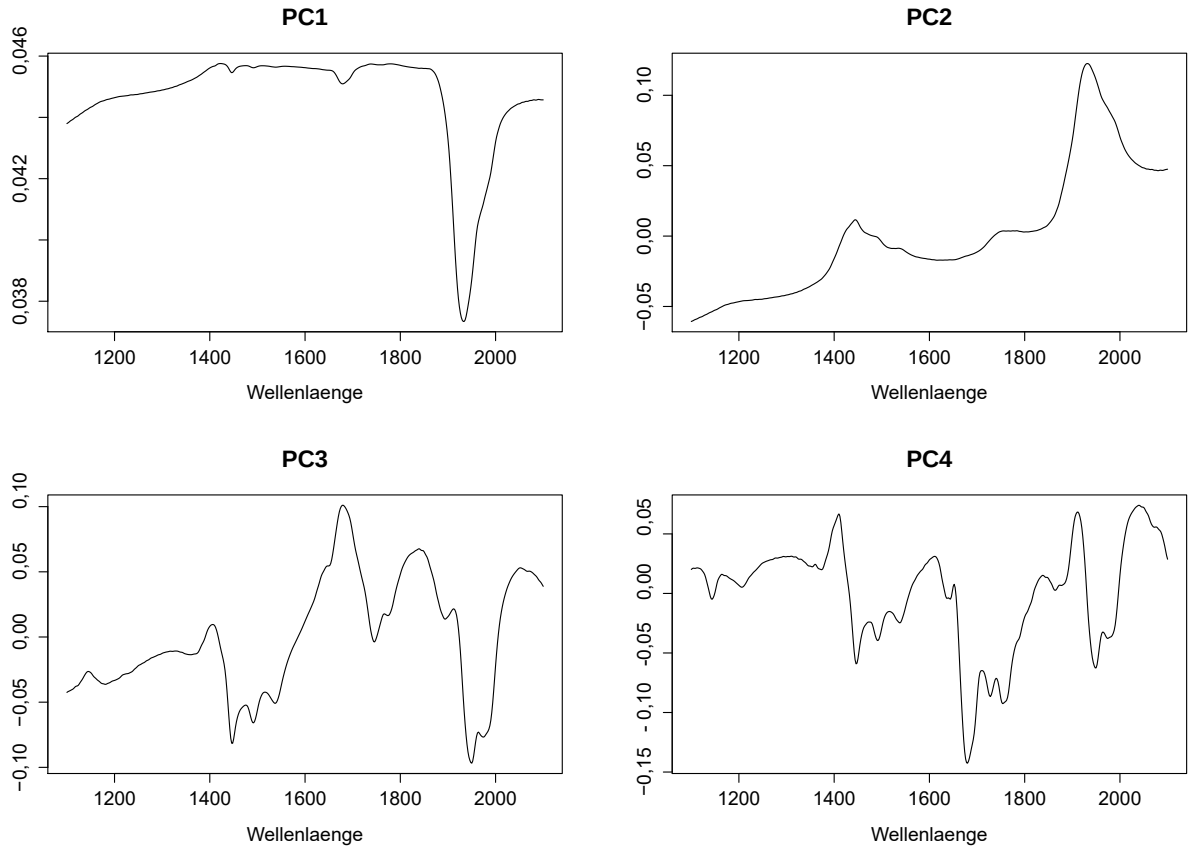

Abbildung 46: Erste vier Hauptkomponenten der IR-Spektren

Die Hauptkomponentenanalyse wurde auf dem Spektraldatensatz nach einer vorherigen Zentrierung und Normierung angewandt. Die ersten Hauptkomponenten (*principal component* - PC) (Abb. 46) zeigen die höchste Variation der IR-Spektren. Die erste Hauptkomponente (PC1) zeigt eine starke Absorptionsbande bei 1920 nm. PC2 zeigt einen Anstieg mit zwei signifikanten Maxima bei 1420 und 1920 nm. PC3 und PC4 haben verschiedene Extrema (Minima und Maxima) bei 1420, 1600, 1700, 1750 und 1920 nm. Diese Extrema weisen relevante Wellenlängen und physikalische Eigenschaften von Baumaterialien auf. In der Abbildung 47 sind die Spektren in Form der ersten drei Hauptkomponenten in einem gemeinsamen Koordinatensystem dargestellt.

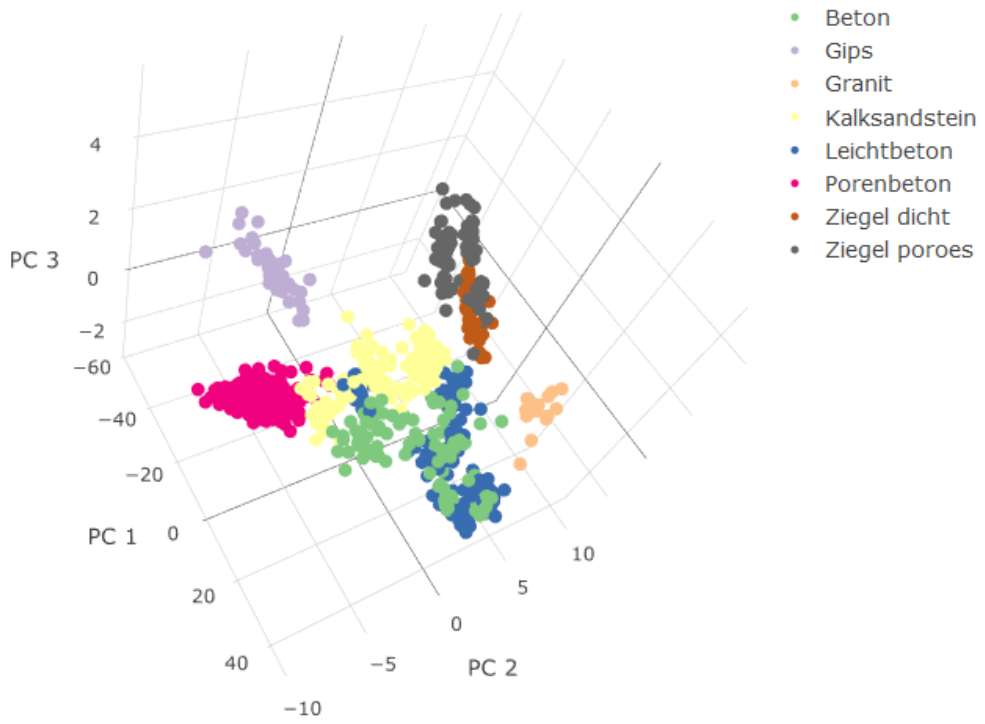

Abbildung 47: Graphische Darstellung von IR-Spektren (8 Materialklassen) in Form von drei Hauptkomponenten

Die Klassen bilden einen großen Hauptcluster, der aus den Klassen Leichtbeton, Beton, Porenbeton und Kalksandstein besteht, und mehrere kleine Lokalcluster für Ziegel, Gips und Granit. Die Klassen Porenbeton und Granit bilden ellipsoidförmige Cluster in Form homogener Punktwolken. Gips formiert einen flächig ellipsoidförmigen Cluster, der in großer Entfernung zum Hauptcluster liegt. Außer ein paar abweichenden Punkten ist dieser Cluster homogen. Die Ziegelklassen bilden zwei naheliegende ellipsoidförmige Cluster, die vom Hauptcluster trennbar sind. Auf der Grenze des Hauptclusters liegt kompakt der Cluster Porenbeton, der eine geringe Überlappung mit dem Cluster Kalksandstein hat. Hingegen haben die Klassen Beton, Leichtbeton und Kalksandstein mehrere Überlappungen und Durchdringungen (insbesondere Beton und Leichtbeton). Einen inhomogenen Cluster bildet die Klasse Beton, die die Klassen Leichtbeton, Kalksandstein und ein wenig auch die Klasse Porenbeton überlappt, dies verhindert eine gute Trennbarkeit von Beton von diesen Klassen bei der Anwendung von 3 Hauptkomponenten.

Die Abbildung 48 zeigt die Spektren in Form der ersten drei Hauptkomponenten, wenn die Spektren in 5 Oberklassen eingeordnet sind. Es ist ersichtlich, dass die Klasse 4 eine starke Überlappung mit den Klassen 1, 2, 3 hat. Wegen der vereinigten Materialklassen Porenbeton, Leichtbeton und porosierter Ziegel ist es schwieriger, die einzelnen Klasse von einander zu trennen, was zu einer Verringerung der Erkennungsrate führt.

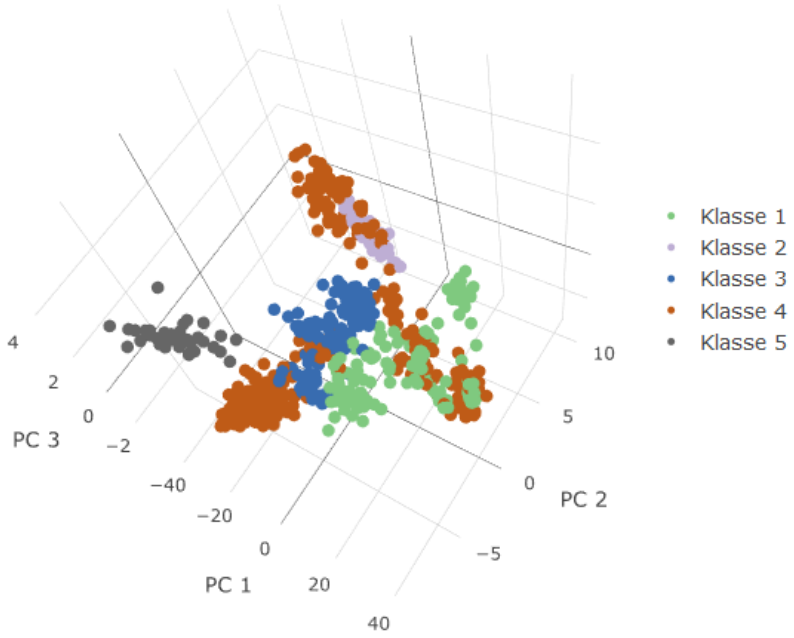

Abbildung 48: Graphische Darstellung von IR-Spektren (5 Oberklassen) in Form von drei Hauptkomponenten

Die Hauptkomponenten wurden dann für das Training und den Test der ausgewählten Klassifikatoren verwendet. Der Einfluss der Komponentenanzahl auf die Erkennungsrate ist in den Abbildungen 49 und 50 dargestellt.

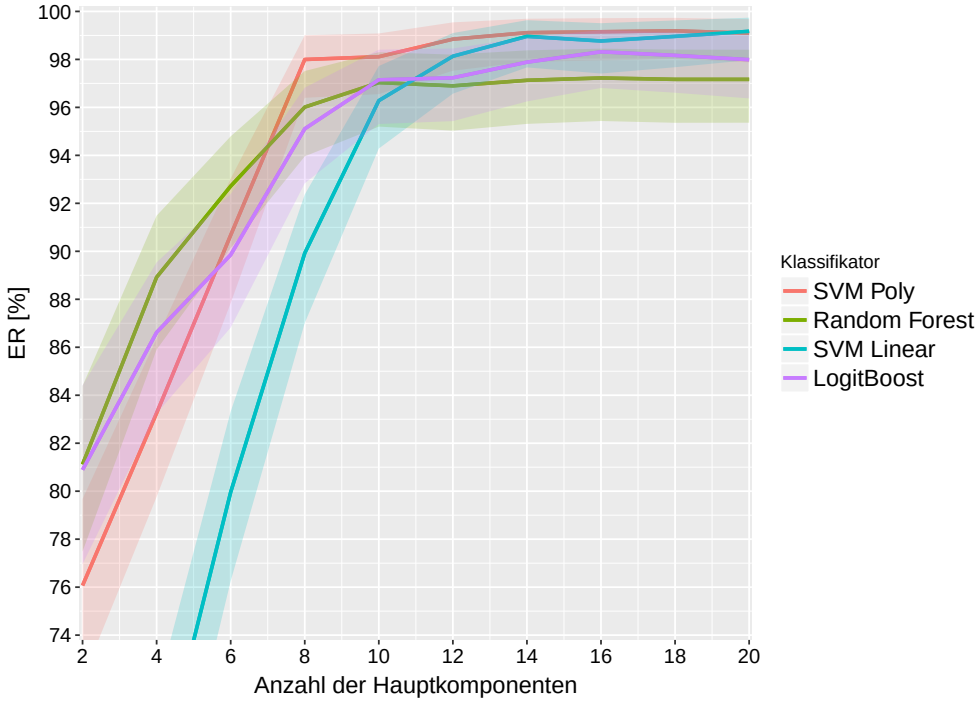

Abbildung 49: Erkennungsraten bei Anwendung der Hauptkomponenten aus den IR-Spektren (5 Oberklassen)

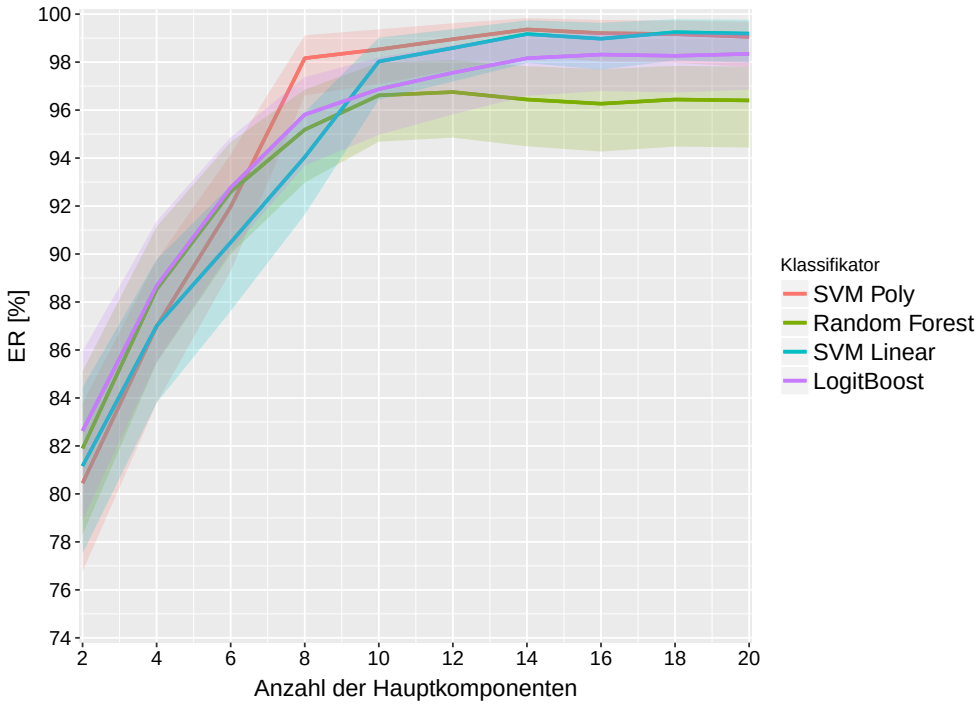

Abbildung 50: Erkennungsraten bei Anwendung der Hauptkomponenten aus den IR-Spektren (8 Materialklassen)

Bei alleiniger Verwendung der ersten zwei Hauptkomponenten ist die Leistung schlecht und die GER ist weniger als 82% für alle Klassifikatoren. Die Anwendung weiterer Hauptkomponenten zeigt einen deutlichen Anstieg der Erkennungsrate bis auf 98% bei der Anwendung von mehr als 14 Hauptkomponenten (außer Random Forest). Die einzelnen Erkennungsraten sind höher als 90% bei Einsatz von 8 bis 10 ersten Hauptkomponenten abhängig vom Klassifikator. Die besten Ergebnisse auf dem Datensatz mit 5 Oberklassen wurden unter Anwendung der Klassifikatoren svmPoly und svmLinear erreicht - die GER entsprechen 99,2% und 98,8% bei Anwendung der 16 Hauptkomponenten. Die einzelnen Erkennungsraten sind in der Tabelle 20 dargestellt.

Tabelle 20: Klassifikatorleistungen auf dem Spektraldatensatz mit 5 Oberklassen unter Anwendung von 16 Hauptkomponenten

	svmPoly		RF		SVM (Linear)		LogitBoost	
	EER [%]	GER [%]	EER [%]	GER [%]	EER [%]	GER [%]	EER [%]	GER [%]
Klasse 1	98,0		95,7		98,0		97,9	
Klasse 2	99,2		97,1		99,5		98,4	
Klasse 3	98,8	99,2	98,1	97,2	97,1	98,8	97,4	98,3
Klasse 4	99,6		97,2		99,5		98,7	
Klasse 5	100,0		100,0		100,0		100,0	

Drei von vier ausgewählten Klassifikatoren haben eine GER von über 98% auf dem

Datensatz mit 8 Materialklassen unter Anwendung von 16 Hauptkomponenten erreicht. Die höchste erreichte GER ist 99,2% unter Anwendung des Klassifikators svmPoly. Sehr nahe beieinanderliegende Resultate zeigt auch der Klassifikator svmLinear - 99%. Die einzelne Erkennungsraten auf dem Datensatz mit 8 Materialklassen sind in der Tabelle 21 dargestellt.

Tabelle 21: Klassifikatorleistungen auf dem Spektraldatensatz mit 8 Materialklassen unter Anwendung von 16 Hauptkomponenten

	svmPoly		RF		SVM (Linear)		LogitBoost	
	EER [%]	GER [%]	EER [%]	GER [%]	EER [%]	GER [%]	EER [%]	GER [%]
Beton	98,8		95,3		97,8		96,4	
Gips	100,0		100,0		100,0		100,0	
Granit	100,0		100,0		98,6		99,2	
Kalksandstein	98,7	99,2	95,2	96,3	99,1	99,0	97,5	98,3
Leichtbeton	98,8		94,9		98,0		97,5	
Porenbeton	99,9		97,5		99,9		99,9	
Ziegel dicht	98,7		98,0		98,9		98,7	
Ziegel porös	100,0		96,8		100,0		99,8	

Bei vorheriger Hauptkomponentenanalyse unter Verwendung der ersten 16 Hauptkomponenten verbessert sich im Gegensatz zur Verwendung aller 501 wellenlängenspezifischen Informationen die Klassifikationsleistung um $> 2\%$ für svmPoly und svmLinear und um $> 9\%$ für RF und LogitBoost.

Die durchgeführten Untersuchungen mit Klassifikatoren zeigen eine Gesamt-Erkennungsrate von 99% unter Verwendung der ersten 16 Hauptkomponenten auf den Spektraldatensätze mit 5 Oberklassen und mit 8 Materialklassen. Die meisten Fehlklassifikationen treten zwischen den Klassen Leichtbeton und Beton auf. Kritisch ist zudem das Auftreten von Fehlklassifikationen zwischen dichtem und porösem Ziegel. Bei der Wiederverwendung des sortenreinen Bauschutts als hochwertig einsetzbarer Baustoff sind jedoch insbesondere diese Fehlklassifikationen kritisch.

13.5. Lineare Diskriminanzanalyse

Eine weitere Methode für die Merkmalsextraktion ist die Lineare Diskriminanzanalyse (LDA). Während die PCA auf die Varianz im Datensatz fokussiert ist, berücksichtigt die LDA mehr die Interklassenvariabilität, was einen anderen Ansatz darstellt und zu einer möglichen Verbesserung der Klassifikationsleistung führen kann. Die Anwendung der LDA auf dem Datensatz mit 5 Oberklassen ergibt 5 lineare Diskriminanzkomponenten (LD), welche dann zum Training und Test der Klassifikatoren genutzt werden. Die besten Leistungen auf dem Datensatz haben die Klassifikatoren svmPoly und svmLinear gezeigt und die GER 96% und 96,2% entsprechend erreicht. Die einzelne Erkennungsraten sind in der Tabelle 22 dargestellt.

Tabelle 22: Klassifikatorleistungen auf dem Spektraldatensatz mit 5 Oberklassen unter Anwendung von 5 linearen Diskriminanzkomponenten

	svmPoly		RF		SVM (Linear)		LogitBoost	
	E.ER [%]	G.ER [%]	E.ER [%]	G.ER [%]	E.ER [%]	G.ER [%]	E.ER [%]	G.ER [%]
Klasse 1	94,9		85,3		93,6		73,7	
Klasse 2	98,8		77,8		97,7		49,9	
Klasse 3	98,4	96,0	92,2	82,4	98,3	96,2	92,0	73,9
Klasse 5	94,5		85,3		95,7		92,0	
Klasse 6	100,0		44,7		100,0		34,0	

Tabelle 23: Klassifikatorleistungen auf dem Spektrdatensatz mit 8 Materialklassen unter Anwendung von 8 linearen Diskriminanzkomponenten

	svmPoly		RF		SVM (Linear)		LogitBoost	
	E.ER [%]	G.ER [%]	E.ER [%]	G.ER [%]	E.ER [%]	G.ER [%]	E.ER [%]	G.ER [%]
Beton	92,0		83,1		92,3		70,5	
Gips	100,0		83,0		100,0		60,5	
Granit	100,0		91,8		100,0		68,0	
Kalksandstein	97,1	96,4	91,1	86,5	97,5	96,5	91,1	77,3
Leichtbeton	93,6		73,5		94,1		75,3	
Porenbeton	99,6		95,7		99,6		94,4	
Ziegel dicht	98,0		90,9		97,4		74,5	
Ziegel porös	96,7		92,8		95,5		70,3	

Die Anwendung der LDA auf den IR-Spektren der 8 Bauschutt-Klassen ergibt 8 lineare Diskriminanzkomponenten (Abb. 51). Die einzelne Erkennungsraten sind in der Tabelle 23 dargestellt. Unter Anwendung der 8 LD und der Klassifikatoren svmPoly und svmLinear wurden eine GER 96,4% und 96,5% entsprechend erreicht.

Die berechneten Erkennungsraten sind um 2% geringer, als unter Anwendung der Hauptkomponente. Das zeigt eine niedrigere Effizienz der LDA im Vergleich zu PCA auf den Datensätzen.

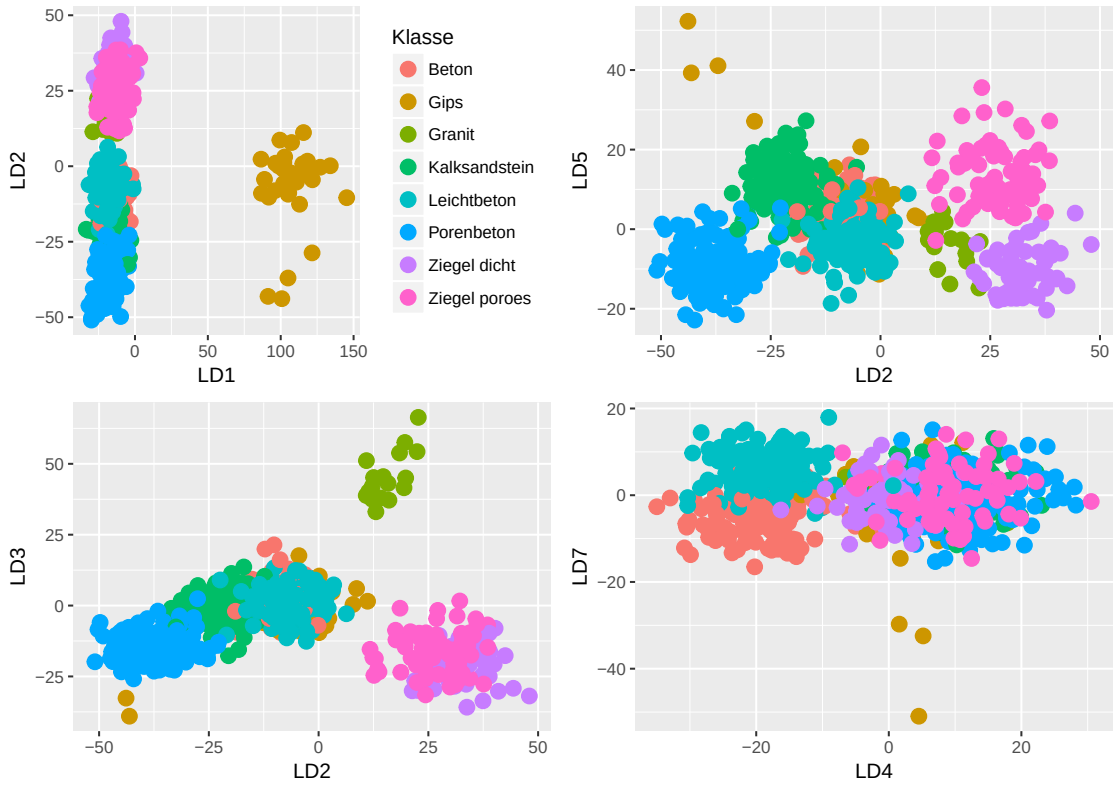

Abbildung 51: Einige Projektionen der linearen Diskriminanzkomponenten

13.6. Merkmalsselektion von Spektren

Bei der Spektralanalyse ist es wichtig, die relevante Information aus der gesamten Datenmenge zu wählen. Die Untersuchungsspektren enthalten mehr als 500 Wellenlängen, die sehr unterschiedlich im Informationsgehalt sind. Um die relevanten Merkmale für die Lösung des Klassifikationsproblems zu finden, müssen die Selektionsverfahren angewendet werden. Die einfachsten und schnellsten Merkmalsselektionsverfahren sind Filter-Verfahren. Diese wurden zuerst wie bei der Bildanalyse getestet. Es wurden die Filter *InfoGain*, *chiSquare-Ranking* und *ReliefF* auf den Datensätze mit 5 Oberklassen und 8 Materialklassen angewendet. Die Ergebnisse der Analyse mit *Info-Gain* sind in den Abbildungen 52 (für 5 Oberklassen) und 53 (für 8 Materialklassen) aufgezeigt.

Bei der Anwendung weniger als 50 Wellenlängen laut *InfoGain* zeigen die Klassifikatoren auf den beiden Datensätze die GER unter 86%. Auf dem Datensatz mit 5 Oberklasen zeigen die Klassifikatoren Random Forest und LogitBoost einen hohen Anstieg der Erkennungsrate und ab Wellenlänge 50 ändern sich die Leistungen nur um 1%. Im Gegensatz zu den Klassifikatoren erreicht der SVM-Klassifikator svmPoly ein Plateau ab 230 Merkmale und zeigt die GER von 97,6% unter Anwendung von 255 Merkmalen. Der andere SVM-Klassifikator svmLinear zeigt einen stetigen Anstieg und hat die besten Resultate bei der Anwendung des vollen Merkmalssatzes. Auf dem Datensatz mit 8 Materialklassen verhalten sich die Klassifikatoren ähnlich, aber der svmLinear-Klassifikator zeigt bessere Ergebnisse und erreicht ein Plateau ab 450 Merkmale.

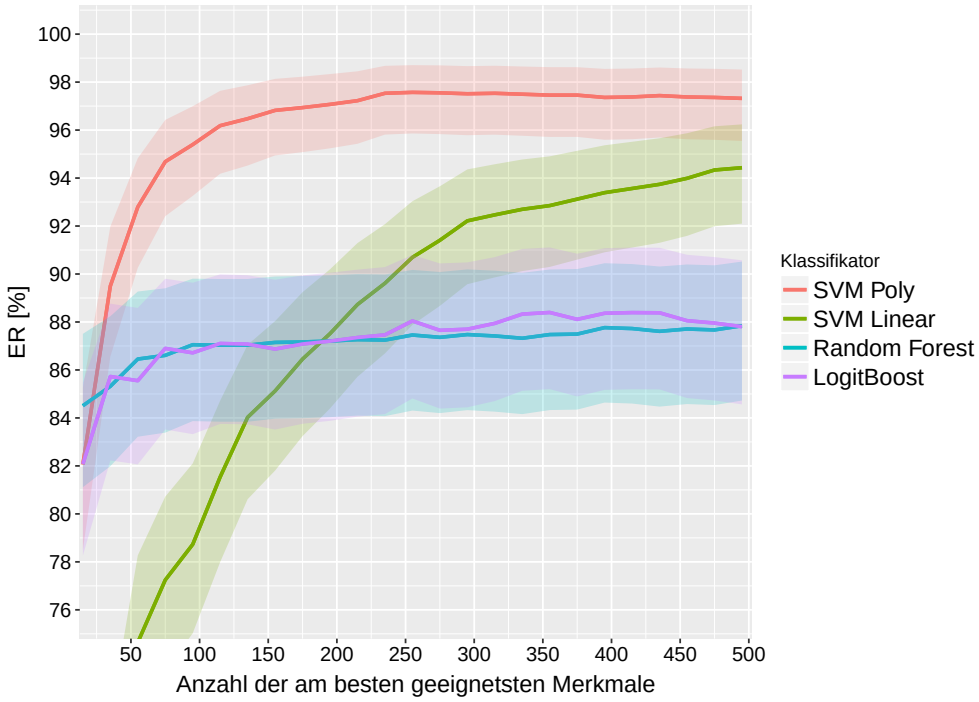

Abbildung 52: Leistungen der Klassifikatoren auf der laut *InfoGain*-Filter besten Wellenlänge vom Datensatz mit 5 Oberklassen

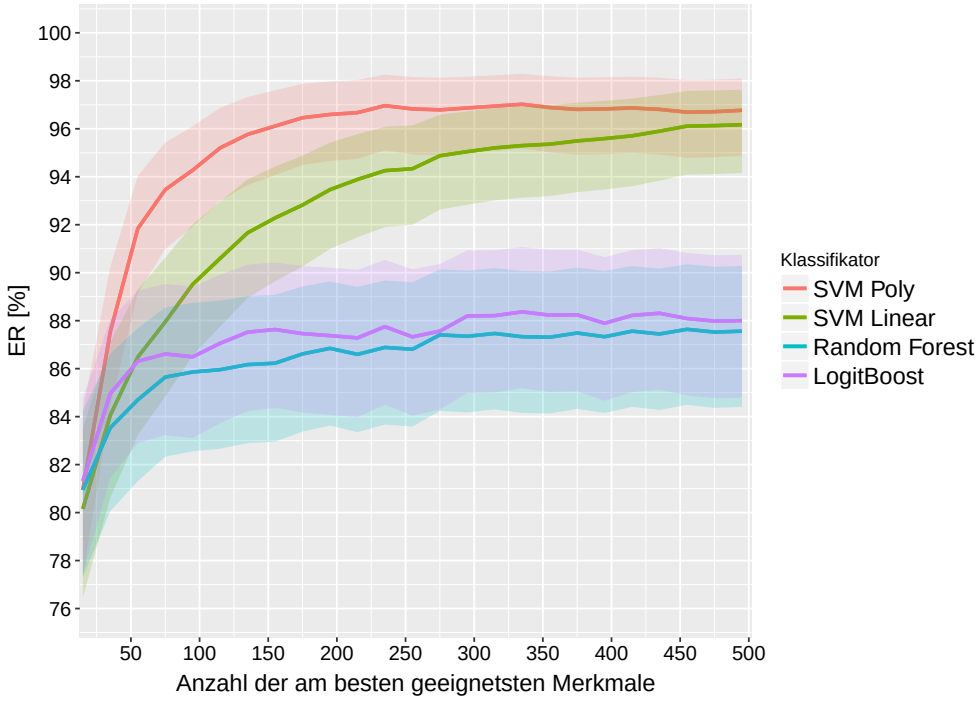

Abbildung 53: Leistungen der Klassifikatoren auf der laut *InfoGain*-Filter besten Wellenlänge vom Datensatz mit 8 Materialklassen

Eine andere getestete Filter-Methode ist das *chiSquare-Ranking*. Die Ergebnisse der Analyse

mit diesem Filter sind in den Abbildungen 54 und 55 dargestellt.

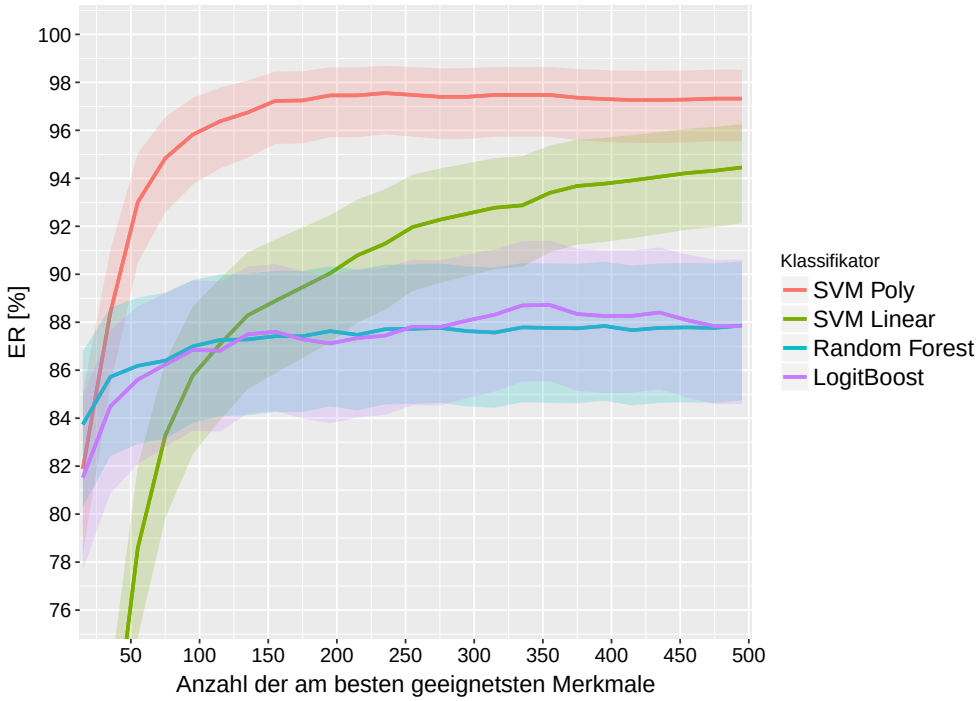

Abbildung 54: Leistungen der Klassifikatoren auf der laut *chiSquare*-Filter besten Wellenlänge vom Datensatz mit 5 Oberklassen

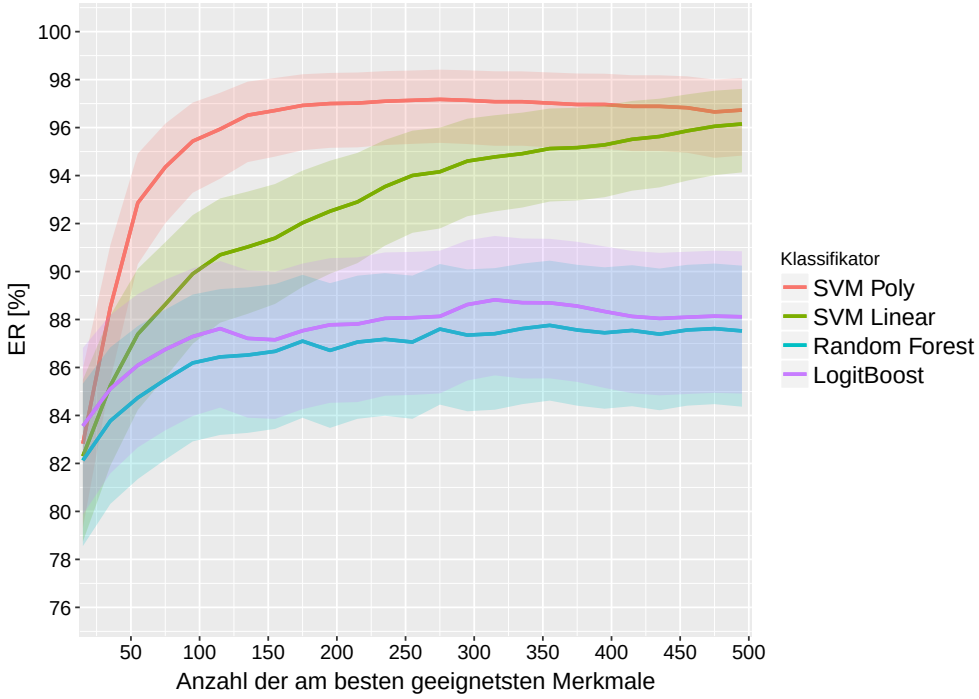

Abbildung 55: Leistungen der Klassifikatoren auf der laut *chiSquare*-Filter besten Wellenlänge vom Datensatz mit 8 Materialklassen

Die Anwendung des *chiSquare*-Filters erlaubt, etwas höhere Gesamterkennungsraten auf dem Datensatz mit 5 Materialklassen unter Anwendung des Klassifikators svmLinear im Bereich von bis zu 450 Merkmale im Vergleich zum InfoGain-Filter zu erreichen, aber die anderen Klassifikatoren zeigen ähnliche Leistungen. Auf dem Datensatz mit 8 Materialklassen zeigen die Klassifikatoren Random Forest (87,8% gegen 87,6%) und LogitBoost (88,8% gegen 88,4%), sowie der Klassifikator svmPoly höhere Erkennungsraten - 97,2% (unter Anwendung von 275 Merkmalen) im Gegensatz zu 97% unter Anwendung des InfoGain-Filters. Der Klassifikator svmLinear zeigt etwas schlechtere Ergebnisse unter Anwendung eines nicht vollständigen Merkmalssatzes laut chiSquare-Filter.

Ein weiteres getestetes Filterverfahren war der ReliefF-Filter. Die Ergebnisse der Anwendung auf den Datensätze mit 5 Oberklassen und 8 Materialklassen sind in den Abbildungen 56 und 57 entsprechend dargestellt. Auf dem Datensatz mit 5 Oberklassen zeigen alle Klassifikatoren eine Verschlechterung der Ergebnissen im Vergleich zur InfoGain-Methode. Auf dem Datensatz mit 8 Materialklassen zeigt nur der Klassifikator LogitBoost eine Verbesserung der Erkennungsrate im Vergleich zum InfoGain-Filter - 89% im Gegensatz zu 88,8%, alle andere Klassifikatoren in Kombination mit dem ReliefF-Filter zeigen schlechtere Leistungen als in Kombination mit dem InfoGain-Filter.

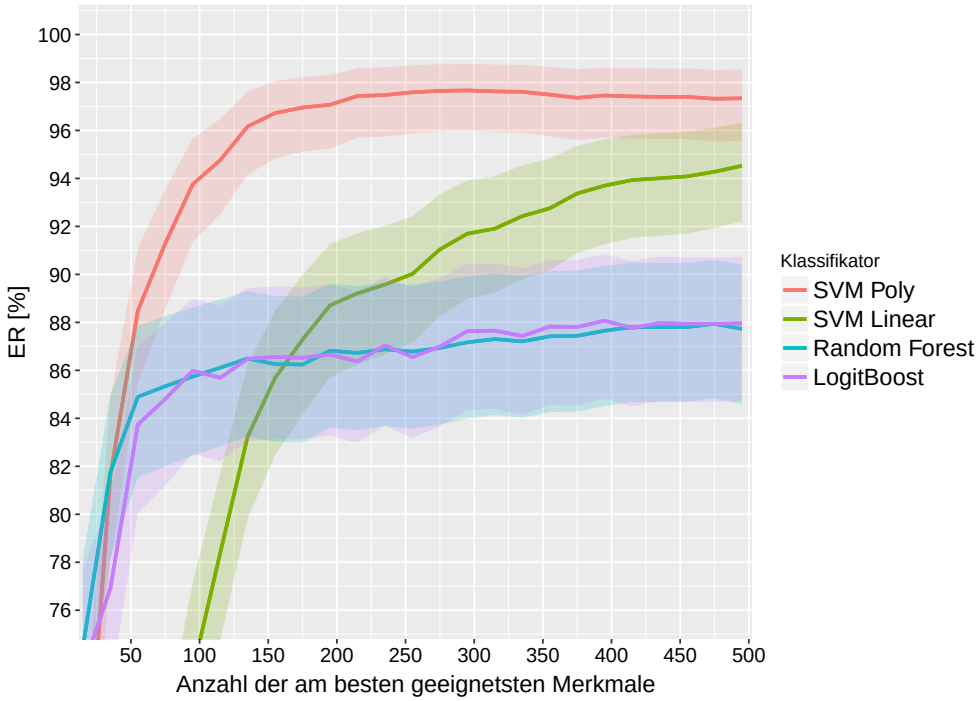

Abbildung 56: Leistungen der Klassifikatoren auf der laut *ReliefF*-Filter besten Wellenlänge vom Datensatz mit 5 Oberklassen

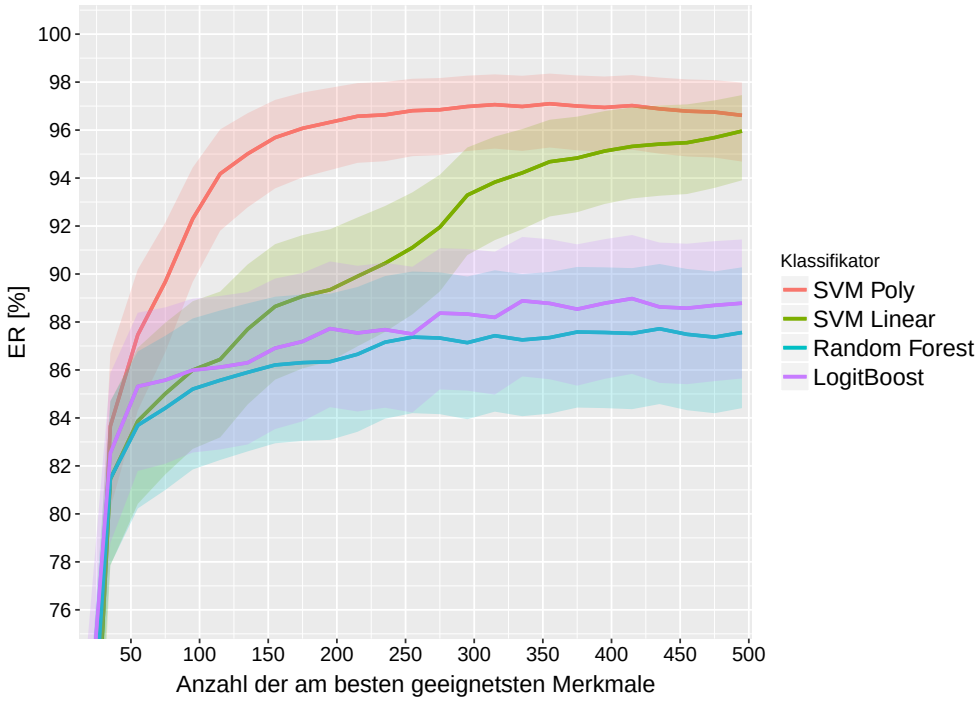

Abbildung 57: Leistungen der Klassifikatoren auf der laut *ReliefF*-Filter besten Wellenlänge vom Datensatz mit 8 Materialklassen

Die durchgeführten Untersuchungen haben gezeigt, dass die optimale Merkmalsanzahl, bei welcher höchsten Leistungen erreicht werden können, im Bereich von 250 bis 300 Merkmale liegt . Der beste Selektionsalgorithmus für die Aufgabe ist InfoGain-Filter.

Außer der PCA, LDA und verschiedenen Filterverfahren wurden für die Merkmalsselektion auch *Wrapper*-Methoden getestet. Diese Verfahren nutzen den Klassifikator selbst für die Merkmalsselektion. Der Vorteil des Verfahrens ist der direkte Zusammenhang der Merkmalsselektion mit dem Klassifikationsprozess, was zum Anstieg der Erkennungsrate führt, aber der Nachteil liegt im deutlich höheren Rechenaufwand.

Für die Untersuchungen mit Wrapper-Methoden wurden zwei Algorithmen ausgewählt:

- *Simulated annealing*, welcher einen Kompromiss zwischen Rechenaufwand und Leistungen darstellt
- *Evolutionärer Algorithmus*, welcher rechenintensiv ist, aber potentiell bessere Leistungen zeigen kann

Es wurden verschiedene Anzahlen von Iterationen für *Simulated annealing* getestet und und der Einfluss auf die Leistungen untersucht. Die optimale Anzahl der Iterationen liegt bei 50. Die Anzahl der Iterationen für den evolutionären Algorithmus wurde wegen des großen Rechenaufwandes auf 50 fixiert. Die Optimierung für die einzelnen Klassifikator hat mehrere Tage gedauert. Die Ergebnisse der Anwendung von *Simulated annealing* und evolutionären Algorithmus auf den Datensätze mit 5 Oberklassen und 8 Materialklassen und Vergleich mit den anderen Merkmalsselektionsalgorithmen sind in der Abb. 58 und 59 dargestellt.

Die simulierte Abkühlung im Gegensatz zum genetischen Algorithmus neigt zur Auswahl von kleineren Merkmalssätzen und zeigt bessere Leistungen unter Anwendung von svmPoly, rf

und LogitBoost-Klassifikatoren. Von allen getesteten Algorithmen haben simulierte Abkühlung und InfoGain-Filter die besten Ergebnisse gezeigt - die simulierte Abkühlung stellt einen Kompromiss zwischen der Anzahl der Merkmale und den Leistungen dar, außerdem hat sie eine relativ geringe Rechenzeit; während der InfoGain-Filter ein stabiles Verhalten für alle verwendeten Klassifikatoren zeigt.

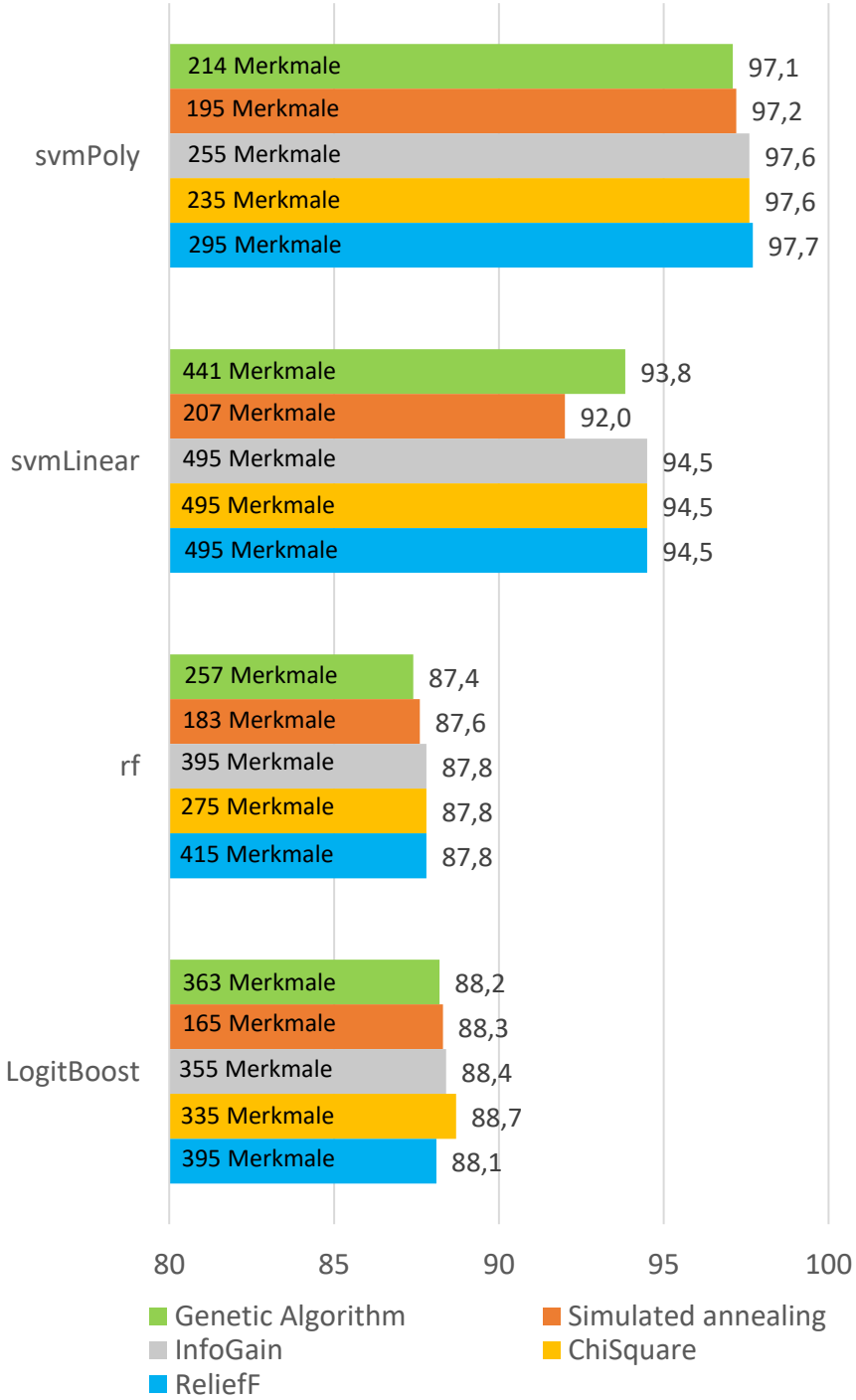

Abbildung 58: Leistungen der Klassifikatoren in Kombination mit verschiedenen Merkmalsselektionsalgortihmen auf dem Datensatz mit 5 Oberklassen

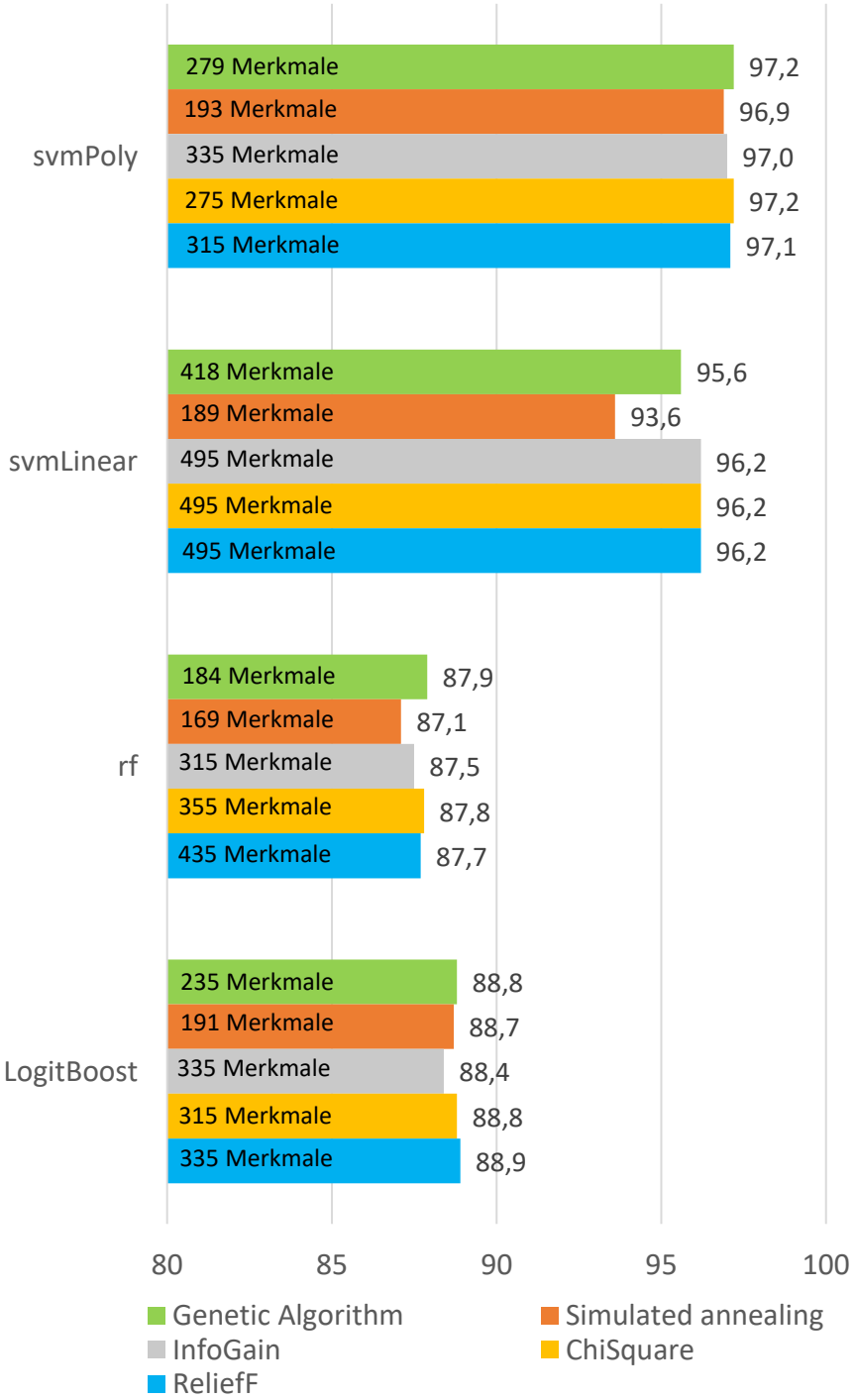

Abbildung 59: Leistungen der Klassifikatoren in Kombination mit verschiedenen Merkmalsselektionsalgortihmen auf dem Datensatz mit 8 Materialklassen

Die getesteten Merkmalsselektionsalgorithmen haben die Leistungen der Klassifikatoren im Vergleich zum originalen Merkmalssatz verbessert - auf dem Datensatz mit 5 Oberklassen: 97,6% gegenüber 96% für svmPoly, 87,8% gegenüber 82,4% für RF, 88,7% gegenüber 73,9% für LogitBoost und keine Änderungen für svmLinear wegen der Anwendung des ganzen Merkmalssatzes; auf dem Datensatz mit 8 Materialklassen: 97,2% gegenüber 96,4% für

svmPoly, 87,9% gegenüber 86,5% für RF, 88,7% gegenüber 77,3% und keine Änderungen für svmLinear wegen der Anwendung vom ganzen Merkmalssatz.

Obwohl die getesteten Merkmalsselektionsalgorithmen eine Verbesserung der Leistungen gezeigt haben, zeigen die beste Ergebnisse die Klassifikatoren auf dem spektralen Datensatz unter Anwendung der Hauptkomponentenanalyse.

13.7. Vergleich der Klassifikationsperformance bei Anwendung des Bild- und Spektraldatensatzes

Der Klassifikator MLP hat unwesentlich bessere Ergebnisse im Vergleich zu den anderen angewendeten Klassifikatoren auf dem Bilddatensatz und eine der schlechtesten Leistungen auf dem Spektraldatensatz gezeigt. Der Klassifikator svmPoly im Gegensatz dazu hat die zweitbesten Leistungen auf dem Bilddatensatz (nur um 0,3-0,4% geringer als MLP) und die besten Leistungen auf dem Spektraldatensatz. Die zwei Klassifikatoren wurden für den Vergleich der Relevanz der Bild- und Spektralinformationen für die Datensätze mit 5 Oberklassen (Tabelle 24) und mit 8 Materialklassen (Tabelle 25) genutzt.

Tabelle 24: Vergleich der besten Leistungen auf den Bild- und Spektraldatensätzen mit 5 Oberklassen

	Bilddatensatz (MLP auf 95 ausgewählten Merkmalen unter Anwendung von ReliefF-Filter)		Spektraldatensatz (svmPoly auf den 16 Hauptkomponenten)	
	EER [%]	GER [%]	EER [%]	GER [%]
Klasse 1	95,3		98,0	
Klasse 2	98,3		99,2	
Klasse 3	95,8	96,9	98,8	99,2
Klasse 4	98,0		99,6	
Klasse 5	94,4		100,0	

Die Gesamterkennungsrate auf den beiden Datensätzen ist bei der Anwendung von spektraler Information um 2,3% höher als bei der Anwendung der Bildinformation. Bei allen Klassen bringt die Anwendung von Spektralinformation den Gewinn: bei einigen Klassen,wie Klasse 2 und Klasse 4, Porenbeton, Ziegel dicht und porös, liegt der Unterschied bei 1%, bei allen anderen Klassen ist der Unterschied größer (bis 6%).

Der Unterschied in den Leistungen kann durch die unterschiedlichen Datensatzgrößen und Datenkomplexität verursacht werden. Im nächsten Teil werden die Bild- und Spektralinformationen mit ähnlichen Datensatzgrößen und Komplexität fusioniert, um höhere Leistungen zu erreichen. Die einzelnen Informationen werden noch einmal untereinander und mit der fusionierten Information verglichen.

Tabelle 25: Vergleich der besten Leistungen auf den Bild- und Spektraldatensätzen mit 8 Materialklassen

	Bilddatensatz (MLP auf 135 ausgewählten Merkmalen unter Anwendung von ReliefF-Filter)		Spektraldatensatz (svmPoly auf den 16 Hauptkomponenten)	
	EER [%]	GER [%]	EER [%]	GER [%]
Beton	95,0		98,8	
Gips	94,2		100,0	
Granit	96,6		100,0	
Kalksandstein	95,3	96,9	98,7	99,2
Leichtbeton	97,1		98,8	
Porenbeton	99,0		99,9	
Ziegel dicht	98,4		98,7	
Ziegel porös	99,0		100,0	

14. Lösung der Klassifikationsaufgabe mit Hybrid-Datensatz

14.1. Kombination der Bild- und Spektralinformation

Während der Untersuchungen auf den Bilddatensatz und Spektraldatensatz wurden hohe Erkennungsraten (die Gesamterkennungsrate von 96,9% für den Bilddatensatz und von 99,2% für den Spektraldatensatz unter Anwendung des svmPoly-Klassifikators) erreicht. Die Kombination von Bild- und Spektralinformation könnte die Erkennungssicherheit erhöhen. Um der Einfluss der zusätzlichen Information auf die Leistungen festzustellen, wurden die Untersuchungen mit einem Hybriddatensatz durchgeführt.

14.2. Aufbau eines Gesamtdatensatzes von Bauschuttproben

Es wurde ein neuer Datensatz vorbereitet, welcher die Bild- und Spektralinformation umfasst. Die Bildinformation wurde mit entsprechende Spektralinformation verknüpft.

Abbildung 60: Beispielbild einer Objektprobe für die Aufnahme des Hybrid-Datensatzes

Dieser Datensatz wurde wegen der fehlenden Verknüpfung zwischen bereits entwickelte Aufbau und dem Spektrometer durch die manuelle Aufnahme erstellt. Anhand der Untersuchungen mit Bildern von Bauschuttproben wurde festgestellt, dass nur die Farb- und Texturinformation für die Klassifikation von Bauschuttmaterialien wert hat (siehe 12.4). Die Messflecke des Messkopfes des Spektrometers beträgt ca. 5 cm und damit größer als einzelne

Objekte. Dabei muss man berücksichtigen, dass Kamera sowie Spektrometer dieselben Objekte aufnehmen müssen. Weil die Relevanz der Formmerkmale für die Klassifikation sehr gering ist, kann der Aufnahmeprozess wie folgt realisiert werden: die einzelnen Proben wurden aus der Gesamtmenge ausgewählt und dicht in eine Schale gelegt (Abb. 60). Der Einfluss des Hintergrundmaterials auf die Spektren ist in diesem Fall beseitigt. Die Objekte in der Schale wurden dann mit dem Spektrometer und mit der Kamera aufgenommen. Auf diese Weise wurden alle 1041 Porben in den Datensatz aufgenommen. Die Anzahl der Objekte pro Klasse hängt von der Anzahl der Subklassen pro Klasse ab. Für weitere Untersuchungen wurde der Datensatz wie beim vorherigen Datensätze so strukturiert, dass er die Proben in die DIN-Oberklassen bzw. Materialklassen zusammenfasst (siehe Tabelle 26).

Tabelle 26: Struktur des Hybriddatensatzes

Klasseneinteilung nach DIN 4226-100	Probenname	Materialklasse	Anzahl an Bsp-Objekten
Beton und Gesteinskörnung	B1, B2, B3, B4, B5, B6, B7, B8, B9, B10	Beton	149
	Granit	Granit	25
Klinker / nicht porosierter Ziegel (Dichte >2 g/cm^3)	Z4, Z9, Z10	Ziegel dicht	75
Kalksandstein	KS1, KS2, KS3, KS4, KS5, KS6, KS7, KS8	Kalksandstein	199
Andere mineralische Bestandteile	LB1, LB2, LB3, LB4, LB5, LB6, LB7, LB9	Leichtbeton	200
	PB1, PB2, PB3, PB4, PB5, PB6, PB7, PB8, PB9	Porenbeton	225
	Z1, Z2, Z3, Z5, Z6, Z7, Z8	Ziegel porös	100
Fremdbestandteile	Ansetzgips, AnsetzgipsE2, Gips, Gipskarton, Gipskarton1	Gips	50

14.3. Evaluierung geeigneter Klassifikationsverfahren in R

Die Klassifikationsverfahren aus der Bibliothek *caret*, welche bei den erweiterten Tests auf dem Bilddatensatz und Spektraldatensatz angewendet wurden, wurden auch auf dem Hybriddatensatz angewendet. Die folgende Klassifikatoren wurden verwendet:

- *Random Forest (RF)*

- *Logistic Regression (LogitBoost)*
- *Support Vector Machine mit linear Kernel (svmLinear)*
- *Support Vector Machine mit polynominal Kernel (svmPoly)*
- *Multilayer Perceptron (MLP)*

14.3.1. Untersuchung des Einflusses von Datenaufteilung in Trainings- und Testpartitionen

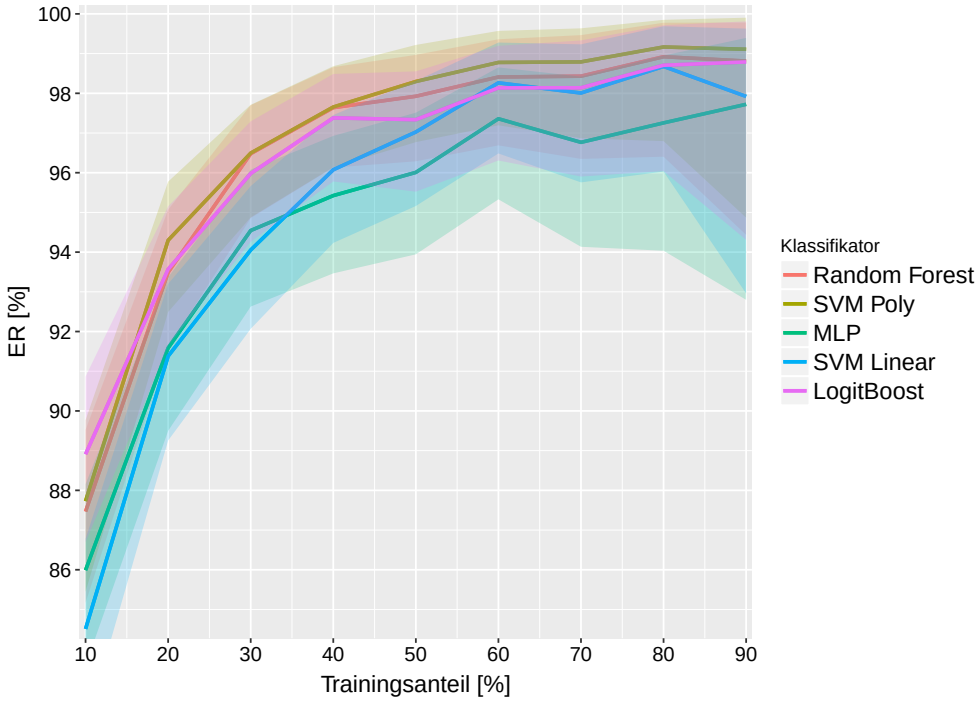

Abbildung 61: Leistung der Klassifikatoren auf dem kleinen Bilddatensatz mit 5 Oberklassen abhängig vom Trainingsanteil

Die Datensatzgröße ist relativ klein. Im Vergleich zum ersten Bilddatensatz enthält der Hybriddatensatz 29 mal weniger Objekte. Deshalb muss ein passendes Validierungsschema wie in dem Fall des spektralen Datensatzes gewählt werden. Das *Hold-out*-Validierungsschema wird wieder getestet. Das Trainingsprozess wie in dem Fall des Spektraldatensatzes wird bei den verschiedenen Trainingsanteilen von 10 bis 90% von der gesamten Datensatzgröße durchgeführt. Jede Trennung wird zehnmal zufällig wiederholt und die Ergebnisse wurden am Ende gemittelt. Die Abbildungen 61 und 62 stellen die Ergebnisse dar.

Aus der Abbildungen ist ersichtlich, dass alle Klassifikatoren große Abweichungen der Erkennungsrate beim Trainingsanteil größer als 60% haben.

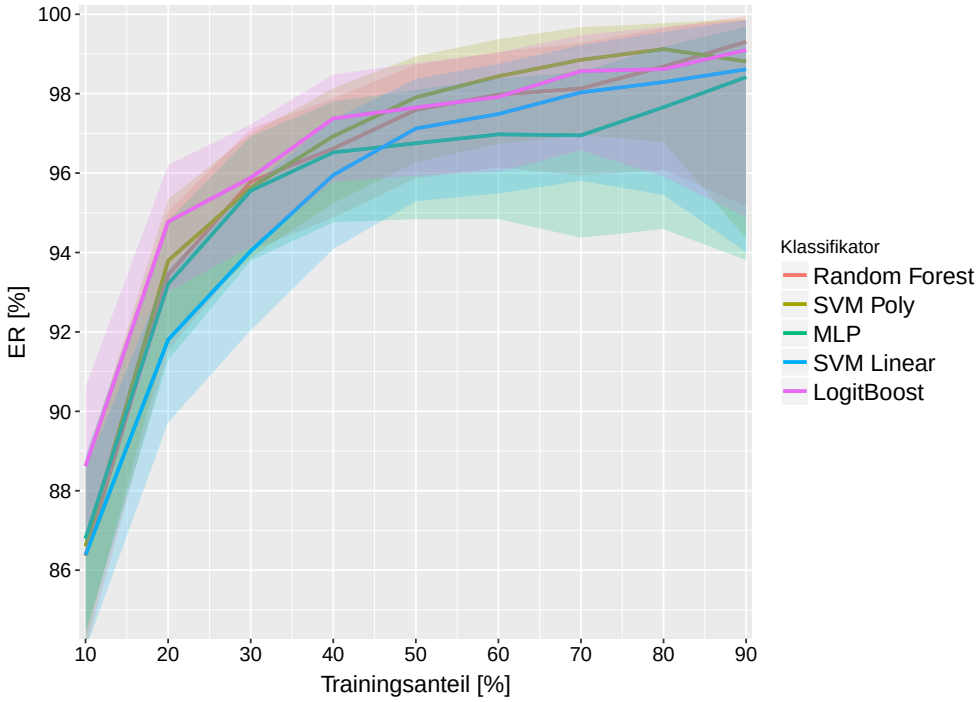

Abbildung 62: Leistung der Klassifikatoren auf dem kleinen Bilddatensatz mit 8 Materialklassen abhängig vom Trainingsanteil

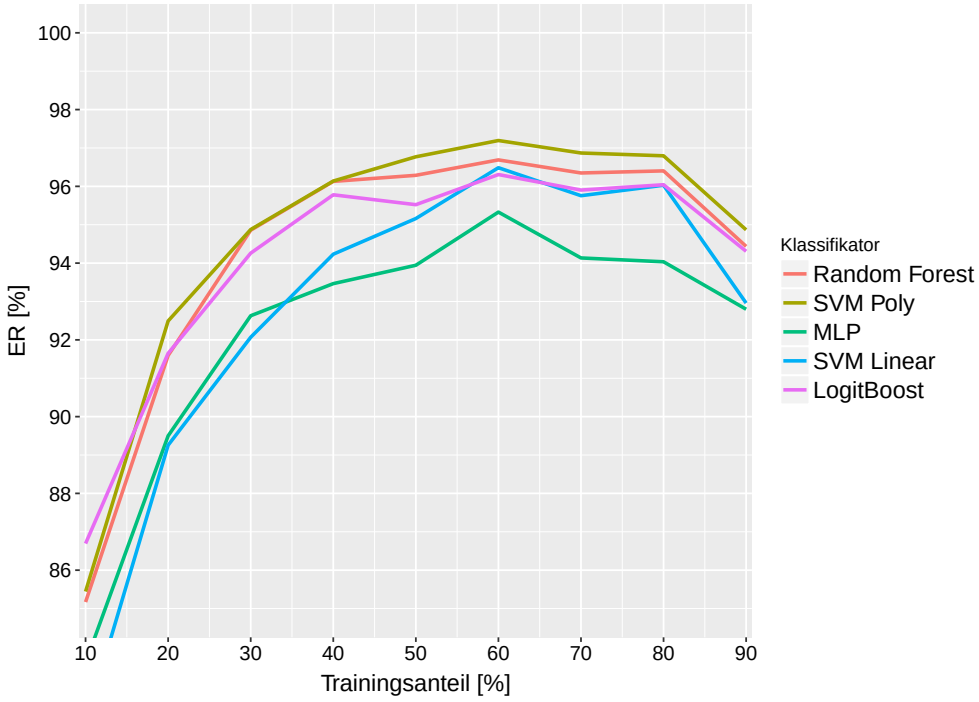

Abbildung 63: Kleinste erreichte Gesamterkennungsraten der Klassifikatoren auf dem kleinen Bilddatensatz mit 5 Oberklassen abhängig vom Trainingsanteil

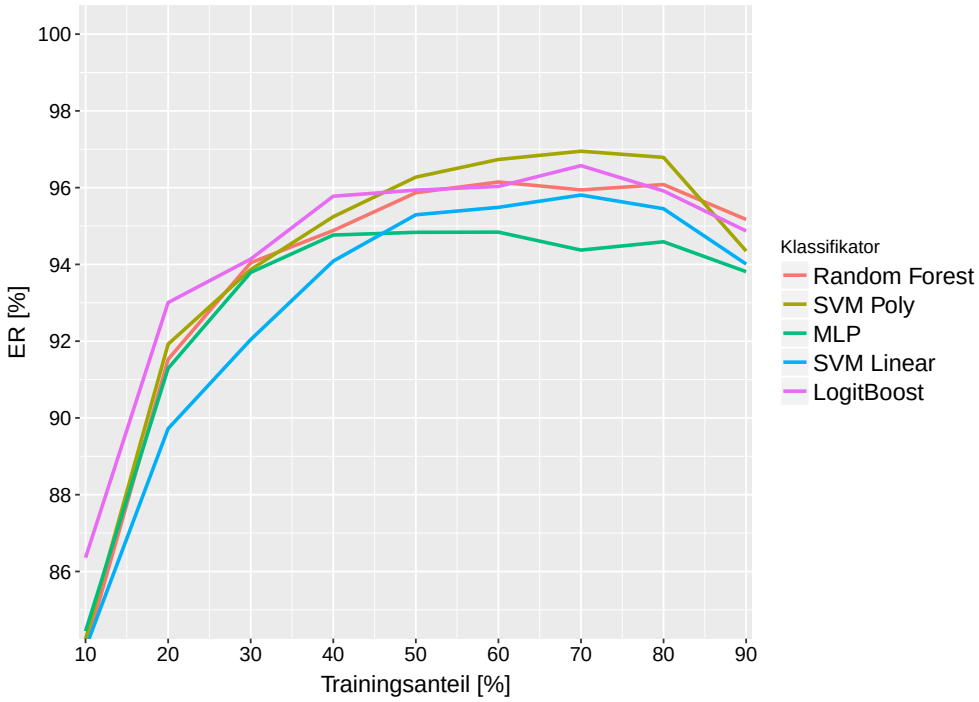

Abbildung 64: Kleinste erreichte Gesamterkennungsraten der Klassifikatoren auf dem kleinen Bilddatensatz mit 8 Materialklassen abhängig vom Trainingsanteil

Die Abbildungen 63 und 64 zeigen die kleinsten erreichten Erkennungsraten. Die Klassifikatoren mit Embedded-Merkmalsselektion Random Forest und LogitBoost erreichen ein Plateau ab 40% auf den beiden Datensätze und weisen wie in dem Fall mit dem Spektraldatensatz mehr Robustheit im Vergleich zu Algorithmen ohne integrierte Merkmalsselektion auf. Andere Klassifikatoren zeigen die Stabilisation und dann den Abfall der Leistungen ab 50-60%. Die vergrößerte Variation der Leistungen verursacht die Überanpassung der Klassifikatoren, weswegen ein optimaler Trainingsanteil bei 50% liegt. Das lässt die Unteranpassung und die Überanpassung beim Training und Test vermeiden. Weil der optimale Trainingsanteil für die Bild- und Spektralinformation des Datensatzes (siehe Kapitel 14.3.1) bei 50% liegt, wurde der Trainingsanteil für die Untersuchungen mit den Hybriddatensätzen auch auf 50% festgelegt.

14.3.2. Anwendung der Klassifikatoren auf dem Hybriddatensatz

Die ausgewählten Klassifikatoren wurden auf den Hybriddatensätzen mit 5 Oberklassen und mit 8 Materialklassen angewendet. Die Ergebnisse sind in den Tabellen 27 und 28 dargestellt.

Die Erkennungsraten aller Klassifikatoren auf dem Hybriddatensatz mit 5 Oberklassen wurden im Vergleich zur alleinigen Anwendung von Spektralinformationen (siehe Tabelle 18) um 1,6% (bei Random Forest) und um 11% (bei MLP) verbessert. Im Gegensatz dazu sind die Leistungen der Klassifikatoren (außer svmLinear und svmPoly) im Vergleich zur alleinigen Anwendung von Bildinformation des Datensatzes (siehe Abb. 61 beim ausgewählten Trainingsanteil 50%) um 0,6% (bei LogitBoost) bis 6,6% (bei Random Forest) schlechter. Auf dem Hybriddatensatz mit 8 Materialklassen haben die Klassifikatoren eine ähnliche Tendenz gezeigt: im Vergleich zur alleinigen Anwendung von Spektralinformation (siehe Tabelle 19) um 0,4% (bei svmPoly) bis 10% (bei MLP) verbessert und im Vergleich zur alleinigen Anwendung

Tabelle 27: Klassifikatorleistungen auf dem Hybriddatensatz mit 5 Oberklassen

	RF		svmLinear		LogitBoost		svmPoly		MLP	
	EER [%]	GER [%]	EER [%]	GER [%]	EER [%]	GER [%]	EER [%]	GER [%]	EER [%]	GER [%]
Klasse 1	86,1		98,0		95,2		97,4		92,4	
Klasse 2	82,0		96,9		100,0		91,6		91,0	
Klasse 3	86,7	89,4	99,4	98,7	99,7	97,9	98,6	97,9	99,4	95,6
Klasse 4	92,2		98,8		97,6		98,7		95,7	
Klasse 5	100,0		99,6		100,0		100,0		100,0	

Tabelle 28: Klassifikatorleistungen auf dem Hybriddatensatz mit 8 Materialklassen

	RF		svmLinear		LogitBoost		svmPoly		MLP	
	EER [%]	GER [%]	EER [%]	GER [%]	EER [%]	GER [%]	EER [%]	GER [%]	EER [%]	GER [%]
Beton	89,4		96,6		95,6		97,2		95,7	
Gips	100,0		99,6		99,6		100,0		100,0	
Granit	100,0		95,6		91,3		95,5		97,7	
Kalksandstein	84,8	89,1	98,4	97,7	99,2	98	98,4	97,2	99,1	96,3
Leichtbeton	83,2		95,1		96,7		94,8		91,8	
Porenbeton	96,8		100,0		99,4		100,0		99,3	
Ziegel dicht	80,8		96,3		98,9		90,6		91,7	
Ziegel porös	94,9		98,9		99,4		98,6		97,7	

von Bildinformation des Datensatzes (siehe Abb. 62 beim ausgewählten Trainingsanteil 50%) um 0,3% (bei LogitBoost) bis 7,7% (bei Random Forest) verschlechtert. Das zeigt, dass der Spektralteil der Datensätze viel redundante Information enthält, was zur Verschlechterung der Leistungen auf den Hybriddatensätzen führt.

Um das Problem zu lösen, wird die Anwendung von Merkmalsselektion- bzw. Transformationsverfahren benötigt. Die Untersuchungen auf dem Spektralteil des Hybriddatensatzes (siehe Kapitel 13.4 und 13.6) haben gezeigt, dass die besten Leistungen auf dem Spektralteil unter Anwendung der Hauptkomponentenanalyse erreicht wurden.

14.3.3. Anwendung der Hauptkomponentenanalyse auf dem Spektralteil des Hybriddatensatzes

Die Hauptkomponentenanalyse wurde auf den Spektralteil der Hybriddatensätze angewendet. Das Ergebnis der Analyse sind die Hauptkomponenten, welche zusammen mit dem unveränderten Bildteil der Datensätze für das Training und den Test der Klassifikatoren verwendet wurden. Die Anzahl der verwendeten Hauptkomponenten wurde von 2 bis 20 variiert. Die Ergebnisse für zwei Hybriddatensätze sind in der Abb. 65 und 66 dargestellt.

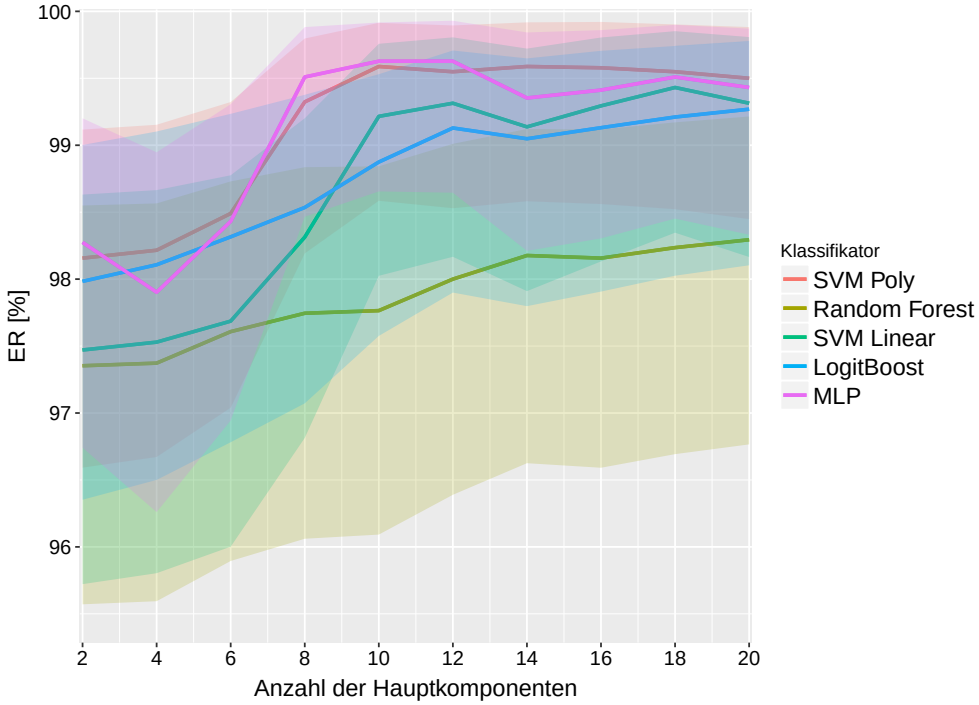

Abbildung 65: Erkennungsraten bei Anwendung der Kombination von Hauptkomponenten aus dem Spektralteil und dem unveränderten Bildteil des Hybriddatensatzes (5 Oberklassen)

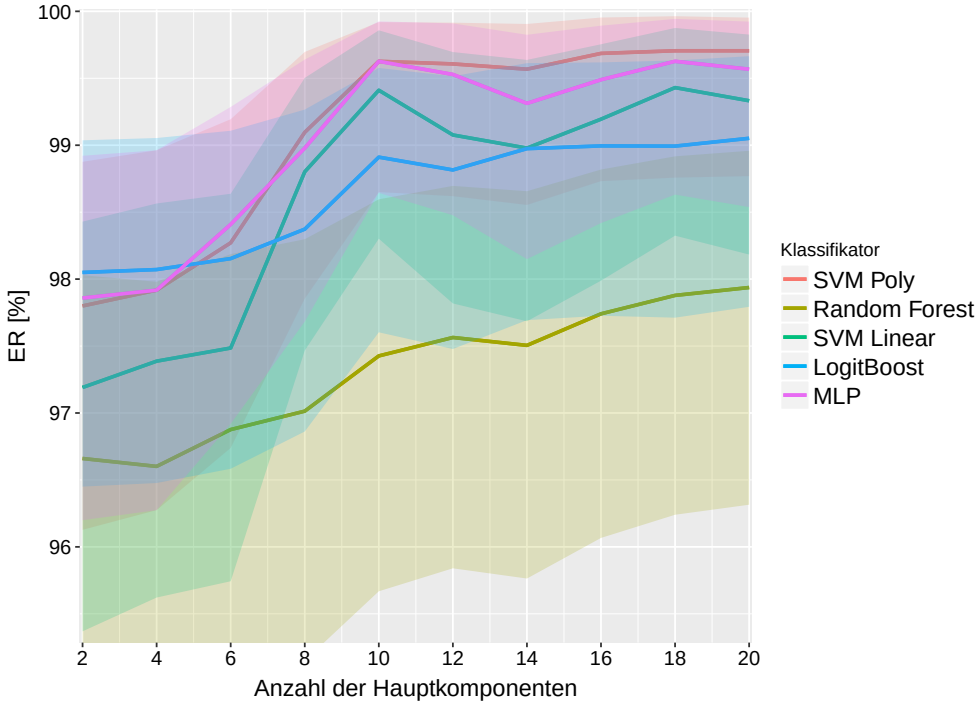

Abbildung 66: Erkennungsraten bei Anwendung der Kombination von Hauptkomponenten aus dem Spektralteil und dem unveränderten Bildteil des Hybriddatensatzes (8 Materialklassen)

Tabelle 29: Klassifikatorleistungen auf dem Hybriddatensatz mit 5 Oberklassen unter Anwendung von Hauptkomponentenanalyse

	RF (20 HK)		LogitBoost (20 HK)		svmPoly (14 HK)		svmLinear (18 HK)		MLP (12 HK)	
	EER [%]	GER [%]	EER [%]	GER [%]	EER [%]	GER [%]	EER [%]	GER [%]	EER [%]	GER [%]
Klasse 1	97,8		98,2		99,7		98,9		99,9	
Klasse 2	97,6		100,0		97,4		98,4		98,2	
Klasse 3	98,1	98,3	99,5	99,3	99,9	99,6	100,0	99,4	100,0	99,6
Klasse 4	98,5		99,4		99,7		99,2		99,6	
Klasse 5	100,0		100,0		100,0		100,0		100,0	

Aus der Abbildungen ist ersichtlich, dass die Klassifikatoren svmPoly und MLP die besten Leistungen von allen Klassifikatoren zeigen und die höchste Erkennungsrate unter Anwendung von 12 und 14 Hauptkomponenten entsprechend (auf dem Hybriddatensatz mit 5 Oberklassen) und von 18 Hauptkomponenten (auf dem Hybriddatensatz mit 8 Materialklassen) erreichen. Eine weitere Vergrößerung der Anzahl der Hauptkomponenten führt zur Verschlechterung der Leistungen.

Die besten Leistungen der Klassifikatoren auf den Hybriddatensätze mit 5 Oberklassen und 8 Materialklassen sind in den Tabellen 29 und 30 entsprechend dargestellt. Die Klassifikatoren haben die Leistungen unter Anwendung einer unterschiedlichen Anzahl von der Hauptkomponenten (HK) erreicht, was neben dem Klassifikatorsname notiert ist.

Tabelle 30: Klassifikatorleistungen auf dem Hybriddatensatz mit 8 Materialklassen unter Anwendung von Hauptkomponentenanalyse

	RF (20 HK)		LogitBoost (20 HK)		svmPoly (18 HK)		svmLinear (18 HK)		MLP (18 HK)	
	EER [%]	GER [%]	EER [%]	GER [%]	EER [%]	GER [%]	EER [%]	GER [%]	EER [%]	GER [%]
Beton	97,5		98		99,6		99,1		99,6	
Gips	100,0		100,0		100,0		100,0		100,0	
Granit	94,8		95,0		97,8		97,8		98,5	
Kalksandstein	97,7	97,9	99,8	97,8	100,0	99,7	99,8	99,4	100,0	99,6
Leichtbeton	99,1		98,6		99,6		98,5		98,9	
Porenbeton	97,2		99,6		100,0		100,0		100,0	
Ziegel dicht	97,4		99,4		98,7		99,5		99,2	
Ziegel porös	99,0		99,1		100,0		100,0		100,0	

Die Klassifikatoren svmPoly und MLP haben die besten Leistungen von allen Klassifikatoren auf den beiden Datensätzen gezeigt. Die Algorithmen haben die gleiche Gesamterkennungsrate auf dem Hybriddatensatz mit 5 Oberklassen gezeigt, obwohl sich die einzelnen Erkennungsraten von 0,1 bis 0,8% unterscheiden.

Die höchste Gesamterkennungsrate auf dem Hybriddatensatz mit 8 Materialklassen um

99,7% hat der Klassifikator svmPoly gezeigt und der Klassifikator MLP folgt direkt dahinter mit der GER um 99,6%.

Die durchgeführten Untersuchungen haben die Ergebnisse der Untersuchungen auf den Spektraldatensätzen bestätigt (Kapitel 13.4). Die optimale Anzahl der Hauptkomponenten für die Lösung der Aufgabe liegt bei 14-18 Hauptkomponenten.

Nach der Optimierung des Spektralteil des Hybriddatensatzes gibt es eine weitere Möglichkeit, die Leistungen zu verbessern. Der Bildteil des Hybriddatensatzes wurde in diesen Untersuchungen nicht geändert. Obwohl die Anwendung von Merkmalsselektionsverfahren auf dem Bilddatensatz eine kleine Erhöhung der Leistungen gezeigt hat (siehe Kapitel 12.4), müssen die Verfahren noch mal auf dem Bildteil der Hybriddatensätze angewendet werden.

14.3.4. Anwendung von Merkmalsselektionsverfahren

Die Untersuchungen im Kapitel 12.4 haben gezeigt, dass alle drei Merkmalsselektionsalgorithmen (InfoGain, chiSquare, ReliefF) ähnliche Leistungen zeigen. Für die weitere Untersuchungen mit den Hybriddatensätze wurde das Verfahren InfoGain wegen des geringeren Rechenaufwandes im Vergleich zu zwei anderen verwendet. Die Anzahl der besten Merkmale aus dem Bildteil laut InfoGain-Filter wurde von 20 bis 160 variiert und zusammen mit der optimalen Anzahl der Hauptkomponenten aus dem Spektralteil für Training und Test der Klassifikatoren angewendet. Die Ergebnisse sind in den Abbildungen 67 und 68 dargestellt.

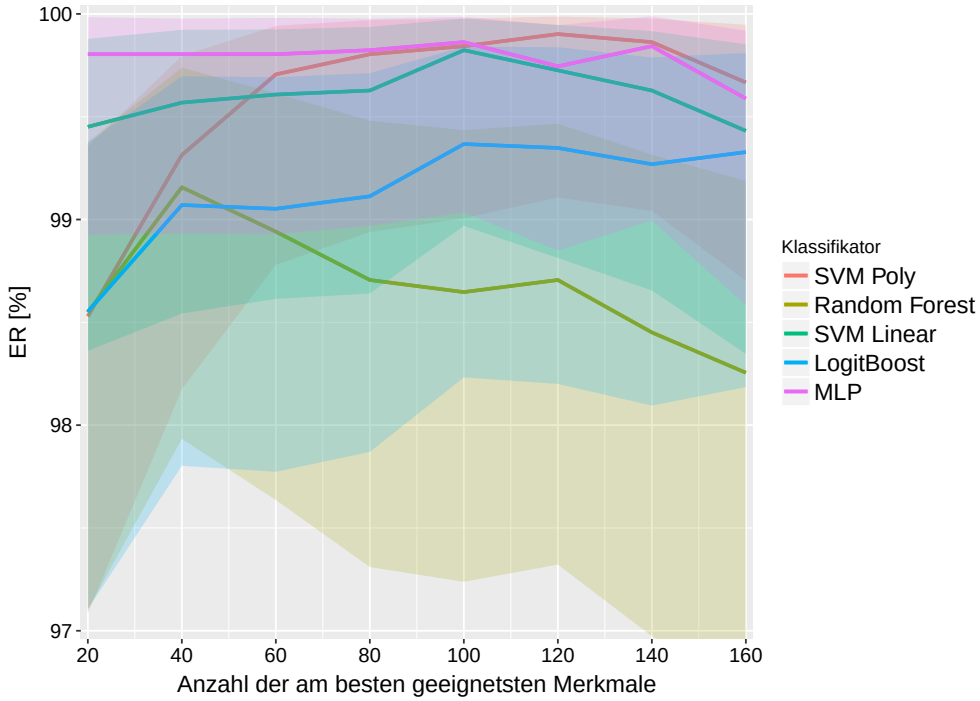

Abbildung 67: Leistungen der Klassifikatoren auf der laut *InfoGain*-Filter besten Merkmale aus dem Bildteil des Hybriddatensatzes mit 5 Oberklassen in Kombination mit Hauptkomponenten aus dem Spektralteil des Datensatzes

Einige Klassifikatoren wie MLP und svmLinear (und LogitBoost auf dem Hybriddatensatz mit 8MK) zeigen nur geringe Schwankungen der Leistungen bei der Änderung der Anzahl

verwendeter Bildmerkmale. Andere Klassifikatoren wie svmPoly und LogitBoost zeigen den Anstieg der Leistungen bei der Vergrößerung der Anzahl von Bildmerkmalen. Der Klassifikator Random Forest zeigt die Verschlechterung der Leistungen bei einer größeren Anzahl der Bildmerkmale.

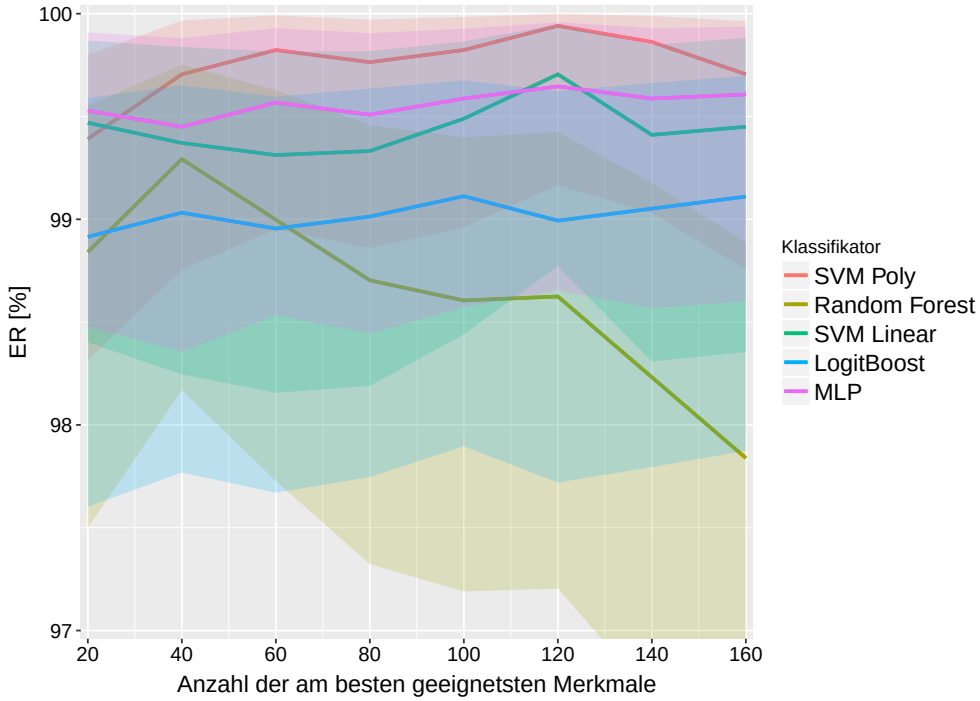

Abbildung 68: Leistungen der Klassifikatoren auf der laut *InfoGain*-Filter besten Merkmale aus dem Bildteil des Hybriddatensatzes mit 8 Materialklassen in Kombination mit Hauptkomponenten aus dem Spektralteil des Datensatzes

Tabelle 31: Klassifikatorleistungen auf dem Hybriddatensatz mit 5 Oberklassen unter Anwendung von InfoGain-Merkmalsselektion in Kombination mit Hauptkomponentenanalyse

	RF (20 HK) (40 Merkmale)		LogitBoost (20 HK) (100 Merkmale)		svmPoly (14 HK) (120 Merkmale)		svmLinear (18 HK) (100 Merkmale)		MLP (12 HK) (100 Merkmale)	
	EER [%]	GER [%]	EER [%]	GER [%]	EER [%]	GER [%]	EER [%]	GER [%]	EER [%]	GER [%]
Klasse 1	99,0		99,2		100,0		99,5		99,9	
Klasse 2	99,5		100,0		99,0		100,0		99,7	
Klasse 3	98,8	99,2	99,6	99,4	100	99,9	100,0	99,8	100,0	99,9
Klasse 4	99,2		99,2		100,0		99,8		99,8	
Klasse 5	100,0		100,0		100,0		100,0		100,0	

Die Anwendung von Merkmalsselektionsverfahren auf dem Bildteil in Kombination mit der Anwendung von Hauptkomponentenanalyse auf dem Spektralteil lässt die höheren Erkennungsraten für alle Klassen bzw. Materialien im Vergleich zur alleinigen Anwendung dieser

Methode zu erreichen. Die Erkennungsraten unter Anwendung von svmPoly-Klassifikator auf dem Datensatz mit 5 Oberklassen sind 100%ig für alle Klassen außer Klasse 2 (99%) (siehe Tabelle 31). Auf dem Datensatz mit 8 Materialklassen zeigt der Klassifikator auch sehr hohe Leistungen und einzelne Materialien, welche nicht mit 100%iger Erkennungsrate detektiert werden können, sind Beton (99,9%) und Leichtbeton (99,8%) (siehe Tabelle 32).

Tabelle 32: Klassifikatorleistungen auf dem Hybriddatensatz mit 8 Materialklassen unter Anwendung von InfoGain-Merkmalsselektion in Kombination mit Hauptkomponentenanalyse

	RF (20 HK) (40 Merkmale)		LogitBoost (20 HK) (100 Merkmale)		svmPoly (18 HK) (120 Merkmale)		svmLinear (18 HK) (120 Merkmale)		MLP (18 HK) (120 Merkmale)	
	EER [%]	GER [%]	EER [%]	GER [%]	EER [%]	GER [%]	EER [%]	GER [%]	EER [%]	GER [%]
Beton	99,1		98,9		99,9		99,7		99,7	
Gips	100,0		100,0		100,0		100,0		100,0	
Granit	100,0		95,2		100,0		100,0		97,0	
Kalksandstein	98,4	99,3	99,6	99,1	100,0	99,9	100,0	99,7	100,0	99,6
Leichtbeton	98,9		98,4		99,8		98,7		98,8	
Porenbeton	99,9		99,6		100,0		100,0		100,0	
Ziegel dicht	100,0		99,7		100,0		100,0		100,0	
Ziegel porös	99,8		99,2		100,0		100,0		100,0	

14.3.5. Vergleich der Relevanz von Bild, Spektral- und Hybridinformationen

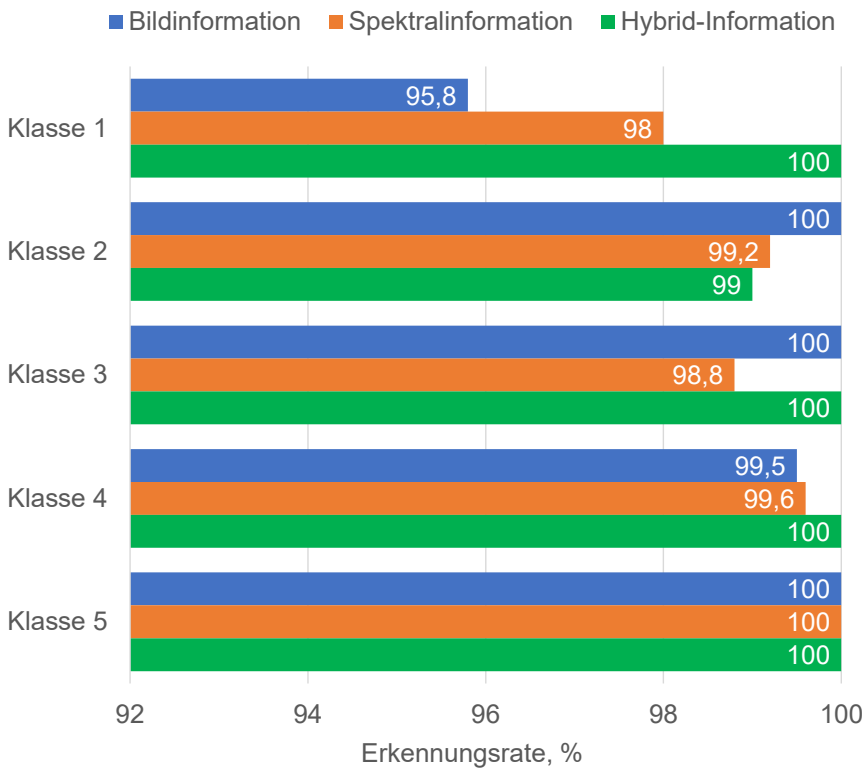

Abbildung 69: Vergleich der Leistungen des Klassifikators *svmPoly* auf den Bild-, Spektral- und Hybriddatensätze mit 5 Oberklassen

Die erreichten Ergebnisse zeigen, dass die Bild und Spektralinformationen zusammen genutzt werden müssen, um die sichere Erkennung von Bauschuttrezyklaten zu realisieren. Obwohl die alleinige Anwendung der Informationen hohe Erkennungsraten (über 95%) erreichen lässt, ist die sichere Erkennung mit ER über 99% nur unter Anwendung von der fusionierten Information möglich (siehe Abb. 69 und 70).

Im Kapitel 13.7 wurde schon bemerkt, dass der Unterschied in Leistungen zwischen verschiedenen Informationen durch unterschiedliche Datensatzgrößen und Datensatzkomplexität verursacht werden kann. Die verwendeten Datensätze in den letzten Untersuchungen haben eine ähnliche Größe und Komplexität, aber der Leistungsgewinn ist für einige Klassen und Materialien nicht sichtbar, weil der Klassifikator die einzelne Erkennungsrate von 100% für die selben Klassen bzw. Materialien auf zwei oder mehr kleinen Datensätzen erreicht. Um die Überanpassung und zu optimistische Leistungen zu vermeiden, wurde der Klassifikator svmPoly auf den Datensätzen mit unterschiedlichen Trainingsanteilen getestet. Die kleineren Trainingsanteile können die Überanpassung vermeiden, aber das kann auch zur Unteranpassung führen.

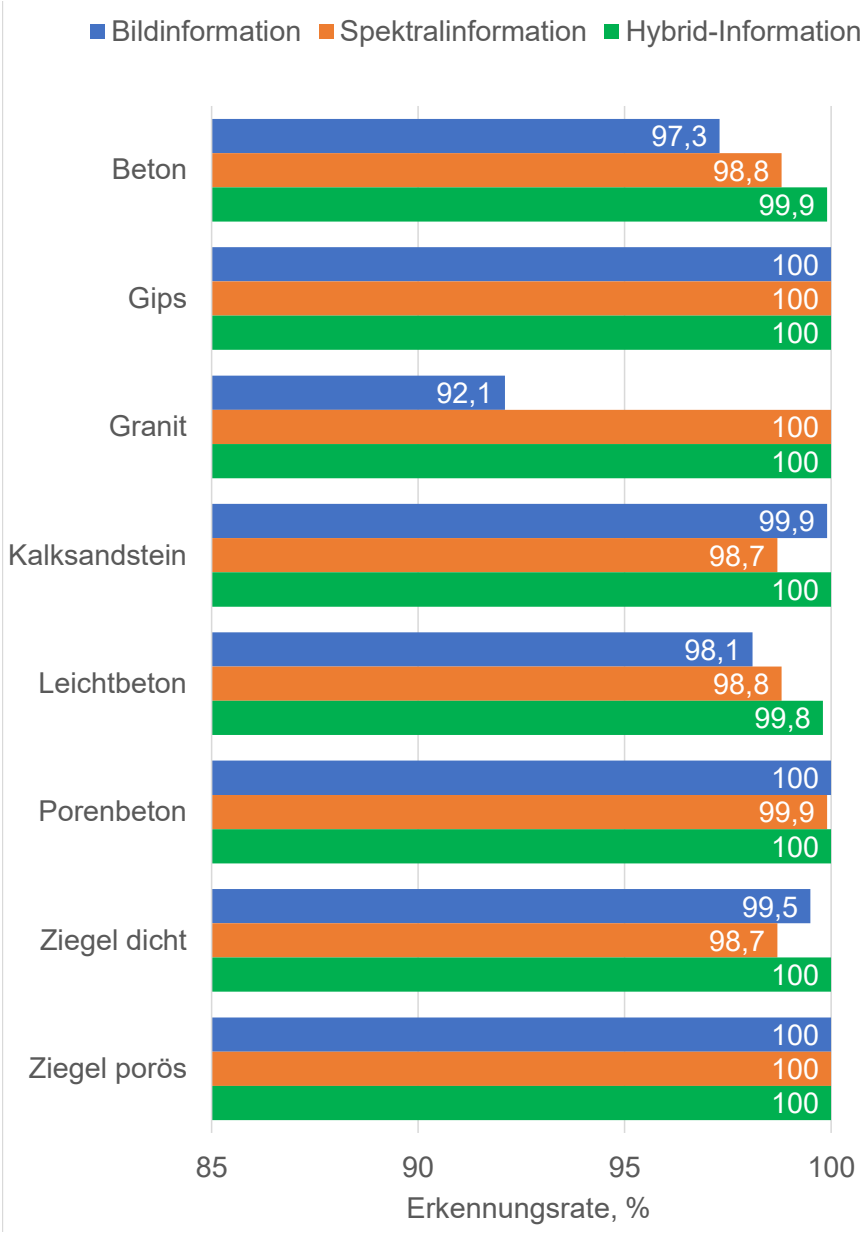

Abbildung 70: Vergleich der Leistungen des Klassifikators *svmPoly* auf dem Bild-, Spektral- und Hybriddatensätze mit 8 Materialklassen

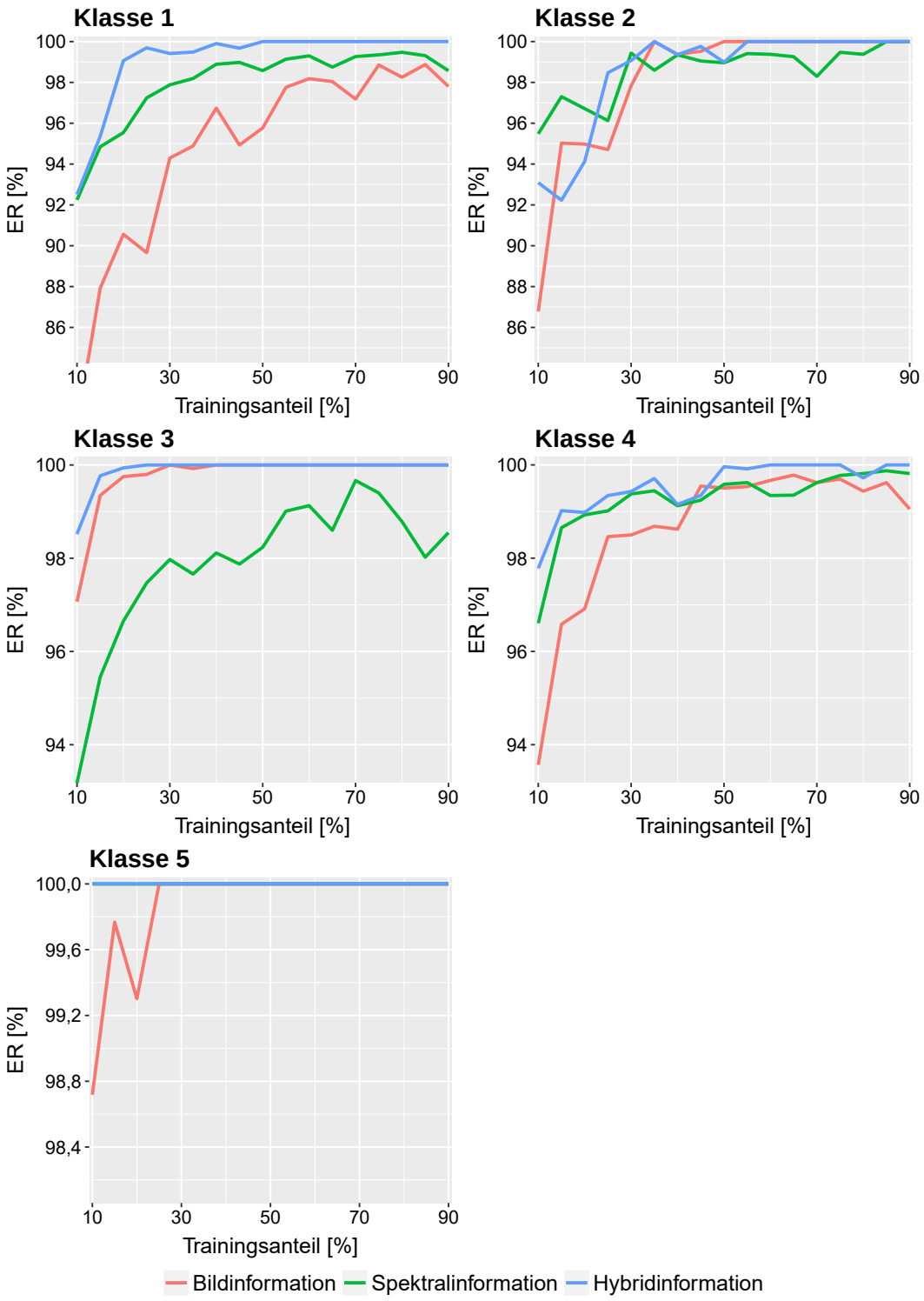

Abbildung 71: Vergleich der Leistungen des Klassifikators *svmPoly* auf dem Bild-, Spektral- und Hybriddatensätze mit 5 Oberklassen abhängig vom Trainingsanteil

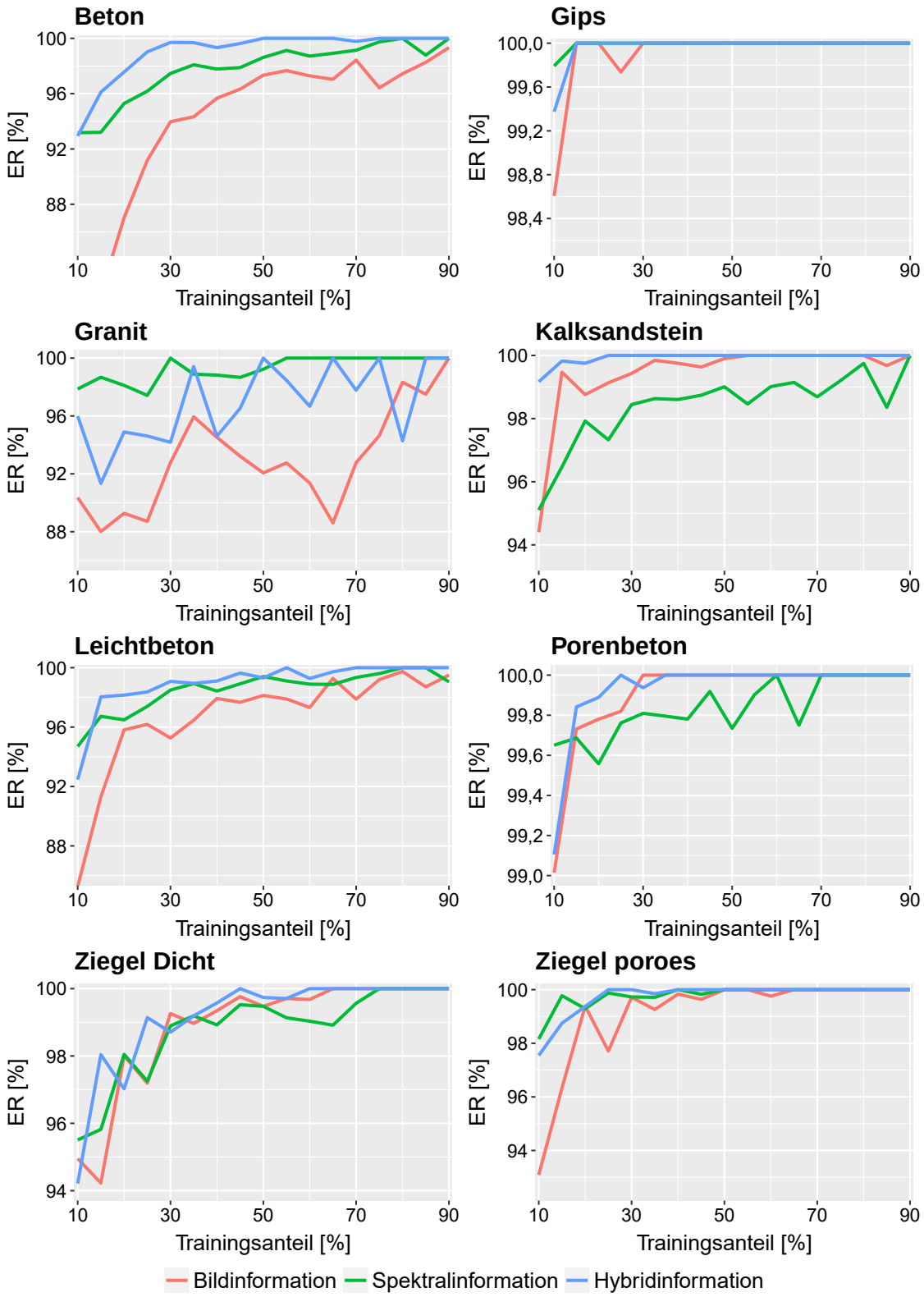

Abbildung 72: Vergleich der Leistungen des Klassifikators *svmPoly* auf dem Bild-, Spektral- und Hybriddatensätze mit 8 Materialklassen abhängig vom Trainingsanteil

Die Abbildungen 71 und 72 zeigen die Testerkennungsraten für einzelne Klassen bzw.

Materialien unter Anwendung des Klassifikators svmPoly abhängig vom Trainingsanteil. Auf jeden Datensatz wurden die am besten geeigneten Merkmalsselektion- bzw. Merkmalsextraktionsverfahren angewendet und die optimale Anzahl der Merkmale gewählt: InfoGain-Merkmalsselektion für die Bilddatensätze, Hauptkomponentenanalyse für die Spektraldatensätze und die Kombination der beiden Algorithmen für die Hybriddatensätze. Für jeden Trainingsanteil wurde die Trennung wie bei den Untersuchungen im Kapitel 13.3.1 zehnmal zufällig wiederholt und die Ergebnisse wurden am Ende gemittelt.

Aus den Abbildungen ist ersichtlich, dass die Anwendung von kombinierter Information den Gewinn im Vergleich zu getrennten Informationen bringt. Die Verwendung der Hybrid-Information ermöglicht eine bessere Erkennung aller Klassen außer der Klasse Granit. Die Erkennung von schwer erkennbaren Klassen, wie Leichtbeton und Beton, wurde um 0,8-2% verbessert. Insgesamt ist das System fähig, die Analyse der Bauschuttrezyklaten mit Erkennungsrate über 99,8% für Materialien und über 99% für Oberklassen durchzuführen. Das System erfüllt die Anforderungen der DIN Norm 4226-100 für die Erkennung von Baustoffkomponenten.

15. Zusammenfassung

Automatisierte Analyse- und Diagnoseverfahren spielen in der heutigen Zeit eine immer größere Rolle in den Bereichen Produktionssicherheit, Lebenswissenschaften und Ressourcenschonung im Rohstoffsektor. Der Grund dafür ist die Erhöhung der Anforderungen in diesen Bereichen bei gleichzeitigem Fortschritt im Bereich der Rechentechnik. Die neuen leistungsfähigen optischen Analyseverfahren erlauben eine schnellere und effektivere Lösung verschiedener Aufgaben in diesen Bereichen. Aktuell zeigt sich insbesondere die Erschließung neuer Rohstofflager bei gleichzeitiger Reduktion des Abfallaufkommens als eine wesentliche Aufgabe. Die Bauindustrie produziert ein großes Abfallvolumen jedes Jahr. Die Abbruchabfälle werden größtenteils (in 2014 zu 81,0 M.-%) im Erd- und Straßenbau verwendet (Downcycling), da hierfür nur geringe technische Anforderungen gelten, und bisher nur ein sehr geringer Teil des anfallenden Bauschutts aufgrund der hohen Heterogenität der Abfälle für die Herstellung von Beton wiederverwendet wird. Bauschuttabfälle stellen Gemische aus mineralischen, organischen und metallischen Bestandteilen dar und können nach einer entsprechenden Aufbereitung als rezyklierte Gesteinskörnungen wiederverwendet werden. Die Qualität der Gesteinskörnungen hat einen großen Einfluss auf die Betoneigenschaften. Die traditionellen Sortiertechniken auf Basis der mechanischen Trennung dominieren bis heute auf dem Recyclingsektor. Diese Verfahren stoßen jedoch an ihre Grenzen bei Stoffen mit ähnlichen Rohdichten. Seit einiger Zeit werden herkömmliche Verfahren in einigen Recyclingsektoren (Papier, Glas, Kunststoff, Holz) bereits durch sensorgestützte Sortierverfahren ersetzt. Die sensorgestützten Verfahren erlauben einen exakteren, schnelleren und preisgünstigeren Sortierprozess. Die Erkennung und Sortierung der Vielzahl an in aufbereiteten Bauabfällen vorkommenden Klassen wurde bisher nur in einigen gezielten Arbeiten untersucht [Linss et al., 2012a], [Anding et al., 2011], [Linss et al., 2017], [Hollstein et al., 2017]. Eine alleinige Anwendung eines Kamerasensors oder alternativ von spektralen Informationen ist laut den Untersuchungen für eine zuverlässige Erkennung unterschiedlicher Komponenten der Bauabfälle nicht ausreichend. Ein modernes optisches System als Kombination von zwei oder mehreren spektralen sowie auch ortsaufgelösten Sensoren könnte unter Verwendung adaptierter Erkennungsverfahren zukünftig in der Lage sein, die Vielzahl der Stoffe im Bauschutt zuverlässig unterscheiden zu können.

Die Anforderungen an Gesteinskörnungen in der Betonherstellung sind im europäischen Standard EN 12620 formuliert, der inhaltlich auf der alten DIN 4226-100 Norm basiert, welche 6 verschiedene Klassen definiert: Beton und rezyklierte Gesteinskörnungen; nicht porosierter Ziegel und Klinker; Kalksandstein; andere mineralische Bestandteile (Leichtbeton, Porenbeton, porosierter Ziegel); Fremdbestandteile (Gips, Glas, Holz, Gipskarton) und Asphalt.

Die bereits am Markt existierenden Geräte sind in der Lage, lediglich einige ausgewählte Komponenten von Bau- und Abbruchabfällen zu sortieren. Trotz großer Vielfältigkeit existiert momentan kein Sortier- oder Analysegerät, welches das komplexe Problem der Erkennung aller Bauschuttkomponenten nach DIN 4226-100 und DIN EN 12620 mit einer hohen Sicherheit lösen kann. Die Automatisierung der Erkennung von Schüttgütern, insbesondere Bauschuttrezyklaten, würde zu einer enormen Zeit- und Kostenersparnis und zur Verbesserung der Ressourceneffizienz führen.

Im Rahmen dieser Arbeit wurde eine Methodik zur automatisierten Analyse von Bauschuttrezyklaten auf der Basis von Bild- und Spektralinformationen entwickelt. Als Basis für die Untersuchungen wurden die Bauschuttproben der Bauhaus Universität Weimar verwendet. Die Proben wurden manuell mithilfe eines Experten in Subklassen nach DIN 4226-100 sortiert.

Sie umfassen 51 Subklassen unterschiedlicher Materialien: 10 Betonproben, 1 Granitprobe, 10 Ziegelproben, 8 Kalksandsteinproben, 8 Leichtbetonproben, 9 Porenbetonproben und 5 Gipsproben. In der Arbeit wurden für die Lösung der Aufgabe die Bild-, Spektral- und Hybrid-Information verwendet. Jede Information wurde ausführlich analysiert und untersucht, was die Anwendung unterschiedlicher Algorithmen aus dem Bereich des maschinellen Lernens nötig machte: relevante Teile der Information mussten ausgewählt, ein optimaler Klassifikator gefunden und für die gegebene Aufgabe optimiert werden. Alle drei Informationen wurden dann mit einander verglichen.

Die ersten Untersuchungen wurden auf dem Bilddatensatz durchgeführt. Mittels des bereits entwickelten Aufbaus des Fachgebietes Qualitätssicherung und industrielle Bildverarbeitung der TU Ilmenau wurde die Bildaufnahme für die Datensatzerstellung realisiert. Dieser Bild-Datensatz enthält mehr als 29 000 Objekte. Für jedes Bild des Datensatzes wurde ein Merkmalsvektor berechnet, welcher 235 Merkmale enthält. Die Merkmale stellen verschiedene Eigenschaften der Objekte dar wie Form, Farbe und Textur. Für die Lösung der Erkennungsaufgabe wurden verschiedene Klassifikatoren verwendet, welche unterschiedliche Klassifikationsansätze darstellen: statistische Klassifikatoren, regelbasierte Klassifikatoren, Entscheidungsbäume, instanzbasierte Klassifikatoren, Support-Vektor-Maschinen, künstliche neuronale Netze. Die zehn folgenden Klassifikatoren wurden getestet: C4.5 Tree, Naive Bayes, k-Nearest Neighbors, Random Forest, Logistic Regression, Support-Vektor-Maschine (SVM) mit linearem, polynomialem und Gauß´schem Kern, Mehrschichtiges Perzeptron (MLP) und Extreme Learning Machine.

Die Klassifikatoren wurden zuerst auf dem Datensatz mit 51 Subklassen angewendet. Die Ergebnisse zeigten, dass komplexe Klassifikatoren wie Random Forest, Logistic Regression, SVM (drei verschiedene Kerne) und MLP höhere Erkennungsraten über 64% im Vergleich zu einfachen Klassifikatoren aufweisen. Die erreichten Erkennungsraten sind jedoch noch nicht ausreichend für die Lösung der Aufgabe. Grund hierfür ist, dass die Datensatzstrukturierung mit 51 Subklassen nicht optimal ist, weil die Subklassen eine geringere Interklassenvariabilität im Vergleich zur Interklassenvariabilität der größeren Oberklassen haben, was zu einer Verschlechterung der Ergebnisse führt. Die Subklassen können unterschiedlich in größere Oberklassen zusammengefasst werden. Am sinnvollsten sehen die Oberklassen nach der DIN 4226-100 Norm wie folgt aus: Beton und rezyklierte Gesteinskörnungen (Klasse 1); nicht porosierter Ziegel und Klinker (Klasse 2); Kalksandstein (Klasse 3); andere mineralische Bestandteile (Leichtbeton, Porenbeton, porosierter Ziegel) (Klasse 4) sowie Fremdbestandteile (Gips, Glas, Holz, Gipskarton) (Klasse 5). Als Alternative wurde auch die Zusammenfassung in 8 Materialklassen durchgeführt: Beton, Granit, Kalksandstein, Leichtbeton, Porenbeton, porosierter und nicht porosierter Ziegel, Gips. Die künstliche Zusammenfassung der 51 Subklassen in 5 Oberklassen nach der DIN 4226-100 Norm und in 8 Materialklassen wurde durchgeführt und ergab für alle Klassifikatoren höhere Erkennungsraten von 71,5 bis zu 96,2%. Obwohl manche Klassifikatoren bei einem Training auf den Subklassen und nachträglichem Zusammenfassung der Ergebnisse eine Gesamt-Erkennungsrate (GER) über 90 % gezeigt haben, ist es rationeller, die Zusammenfassung in größere Gruppen vor dem Klassifikatortraining durchzuführen, damit die Generalisierung erhöht und der Rechenaufwand verringert wird. Für die Untersuchungen auf den Datensätzen mit 5 Oberklassen nach DIN 4226-100 und mit 8 Materialklassen wurden 6 ausgewählte Klassifikatoren verwendet, welche die beste Leistung auf dem Datensatz mit 51 Subklassen gezeigt haben: Random Forest, Logistic Regression, SVM (alle drei verschiedene Kerne) und MLP. Die Erkennungsraten auf

beiden Datensätzen liegen über 93% für alle Klassifikatoren, außer Logistic Regression. Die Datensatzstrukturierung spielt eine geringere Rolle in dieser Situation und der Unterschied in der Leistung zwischen 5 Oberklassen und 8 Materialklassen beträgt 0,1-0,9%. Die besten Ergebnisse erreichten die Klassifikatoren SVM mit polynomialem Kern (GER 96,1% auf dem Datensatz mit 5 Oberklassen und GER 96,3% auf dem Datensatz mit 8 Materialklassen) und MLP (GER 96,5% auf dem Datensatz mit 5 Oberklassen und GER 96,6% auf dem Datensatz mit 8 Materialklassen).

Weitere Verbesserungen der Leistung sind möglich, weil die Merkmalsanzahl in den Untersuchungen sehr groß war und einige Merkmale redundant oder irrelevant und damit kontraproduktiv für die Klassifikation sein können. Zur Optimierung des Merkmalssatzes wurden daher verschiedene Merkmalsselektionsalgorithmen untersucht, welche die unterschiedlichen Prinzipien der Selektion darstellen: Filter- und Wrapper-Verfahren sowie Embedded-Methoden. Die Anwendung von Filterverfahren (InfoGain-Ranking, chiSquare-Ranking und ReliefF-Filter) haben gezeigt, dass die Klassifikatoren sowohl eine Unteranpassung bei einer geringeren Merkmalsanzahl als auch einer Überanpassung bei einer zu großen Merkmalsanzahl aufweisen. Die optimale Merkmalsanzahl liegt für den Klassifikator svmPoly bei 135 Merkmalen und bei 95 für MLP ohne Abhängigkeit von der Art des Filterverfahrens. Laut angewandter Filterverfahren haben Textur- und Farbmerkmale den Haupteinfluss auf die Klassifikationsleistung. Alle untersuchten Filterverfahren (Info-Gain, chiSquare, ReliefF) zeigten ähnliche Ergebnisse auf den beiden Datensätzen. Das InfoGain-Merkmalsselektionsverfahren und das chiSquare-Verfahren hatten den geringsten Rechenaufwand bei gleichzeitig der besten Leistungen auf den beiden Datensätzen, weswegen die Anwendung eines Filter-Selektionsverfahrens für die Klassifikationsaufgabe optimal ist. Die Anwendung der Filterverfahren verbesserte die Gesamterkennungsraten um 0,1-0,5%. Die Anwendung der multivariaten Methode Correlation Feature Selection zeigte hingegen keine Verbesserung.

Eine andere Möglichkeit zur Optimierung des Merkmalssatzes sind Merkmalsreduktionsverfahren wie die Hauptkomponentenanalyse. Die Ergebnisse der Anwendung der Hauptkomponentenanalyse auf den Bilddatensätzen zeigten, dass diese Methode zu einer Verschlechterung der Leistung führte und die Erkennungsaufgabe durch die Einführung komplexerer Merkmale (Hauptkomponenten) erschwert wurde. Laut durchgeführter Untersuchungen ermöglicht bereits die Bildinformation eine hinreichende gute Erkennung folgender Klassen: Klasse 2 (porosierter Ziegel und Klinker) und Klasse 5 (Fremdbestandteile) für den Datensatz mit 5 Oberklassen; Porenbeton und porosierter und dichter Ziegel für den Datensatz mit 8 Materialklassen. Die Schwachstellen stellen hier die Klassen 1, 3 und 5 bzw. Beton, Kalksandstein und Gips dar.

Eine andere Informationsquelle für die optische Bauschutterkennung stellt die Ergänzung spektraler Informationen dar. Hierzu wurden Absorptionsspektren im VIS- und NIR-Bereich aufgenommen. Nur einige Materialien zeigen charakteristische Eigenschaften im VIS-Bereich, die im Folgenden für die Lösung der Klassifikationsaufgabe angewendet werden können, zu nennen sind hier porosierter und dichter Ziegel, Kalksandstein, Porenbeton und Granit. Die anderen Klassen weisen eine schlechte Erkennbarkeit im VIS-Bereich auf. Im NIR-Bereich (1100-2100 nm) weisen die Bauschuttmaterialien mehr Eigenschaften auf, die für eine Trennung relevant sind. Die sechs ausgewählten Klassifikatoren, welche für die Untersuchungen auf den Bilddatensätze verwendet wurden, wurden auch auf den Spektraldatensätzen mit 5 Oberklassen und 8 Materialklassen angewendet. Nur die Klassifikatoren SVM mit linearem und polynomialem Kern zeigten eine GER über 90%. Der Grund liegt in einer hohen Korrelation

zwischen Merkmale des Spektraldatensatzes, weil diese im Prinzip Werte der Spektralkurve darstellen. Eine Anwendung einfacher Klassifikatoren auf dem Spektraldatensatz ist nicht ausreichend für die Lösung der Aufgabe. Weitere Verbesserungen der Leistung sind möglich, wenn der Merkmalssatz optimiert wird. Dafür wurden Merkmalsselektions- und Merkmalsextraktionsverfahren verwendet. Filterverfahren lassen im Vergleich zu Wrapper-Verfahren hier geringfügig höhere Erkennungsraten erreichen. Wrapper-Verfahren ergeben jedoch kleinere Merkmalssätze, was den Vorteil eines reduzierten Rechenaufwandes mit sich bringt. Die getesteten Merkmalsselektionsalgorithmen verbesserten die Leistungen der Klassifikatoren im Vergleich zum originalen Merkmalssatz - auf dem Datensatz mit 5 Oberklassen: 97,6% gegenüber 96% für svmPoly, 87,8% gegenüber 82,4% für RF sowie 88,7% gegenüber 73,9% für LogitBoost; auf dem Datensatz mit 8 Materialklassen: 97,2% gegenüber 96,4% für svmPoly, 87,9% gegenüber 86,5% für RF sowie 88,7% gegenüber 77,3% für LogitBoost.

Außer Merkmalsselektionsmethoden wurden auch Merkmalsextraktionsmethoden, wie die-Hauptkomponentenanalyse und die lineare Diskriminanzanalyse, auf den Spektraldatensätzen untersucht. Die Anwendung der ersten 16 Hauptkomponenten im Merkmalsvektor für die Klassifikation verbesserte die Leistungen der Klassifikatoren um 2% für svmPoly und svmLinear und 9% für RF und LogitBoost. Die durchgeführten Untersuchungen unter Verwendung der Merkmalsextraktions- und Klassifikationsverfahren zeigten eine Gesamt-Erkennungsrate von 99% unter Verwendung der ersten 16 Hauptkomponenten auf den Spektraldatensätzen mit 5 Oberklassen und mit 8 Materialklassen. Das ist deutlich höher als unter Anwendung des ganzen Merkmalssatzes oder des mittels Merkmalsselektionsverfahren reduzierten Merkmalssatzes. Die berechneten Erkennungsraten unter Anwendung der linearen Diskriminanzanalyse sind um 2% geringer als unter Anwendung der Hauptkomponentenanalyse (niedrigere Effizienz der LDA im Vergleich zu PCA). Obwohl die erreichten Erkennungsraten unter Anwendung der Hauptkomponenten bereits hoch sind, treten die meisten Fehlklassifikationen noch zwischen den Klassen Leichtbeton und Beton auf. Kritisch ist zudem das Auftreten von Fehlklassifikationen zwischen dichtem und porösem Ziegel. Bei der Wiederverwendung des sortenreinen Bauschutts als hochwertig einsetzbarer Baustoff sind jedoch insbesondere diese Fehlklassifikationen kritisch. Die Gesamterkennungsrate auf den beiden Datensätzen ist bei der Anwendung von spektraler Information um 2,3% höher als bei der Anwendung der reinen Bildinformation. Bei allen Klassen bringt die Anwendung von Spektralinformation einen Gewinn: bei einigen Klassen, wie Klasse 2 und Klasse 4, Porenbeton, Ziegel dicht und porös, liegt der Unterschied bei 1%, bei allen anderen Klassen ist der Unterschied größer (bis 6%). Um noch höhere Leistungen zu erreichen, wurden daher die Bild- und Spektralinformationen fusioniert. Die Erkennungsraten aller Klassifikatoren auf dem fusionierten Hybriddatensatz mit 5 Oberklassen wurden im Vergleich zur alleinigen Anwendung von Spektralinformation des Datensatzes um 1,6% bei Random Forest bzw. 11% bei MLP verbessert. Im Gegensatz dazu sind die Leistungen der Klassifikatoren (außer svmLinear und svmPoly) im Vergleich zur alleinigen Anwendung von Bildinformation des Datensatzes um 0,6% bei LogitBoost bzw. 6,6% bei Random Forest schlechter. Auf dem Hybriddatensatz mit 8 Materialklassen haben die Klassifikatoren eine ähnliche Tendenz gezeigt. Der Grund dafür ist ein negativer Einfluss der spektralen Merkmale aufgrund ihrer sehr starken Korrelation untereinander. Das Problem kann, ebenso wie beim Spektraldatensatz erfolgt, mittels Hauptkomponentenanalyse gelöst werden. Die Optimierung des Spektralteils des Hybriddatensatzes erlaubte eine Erhöhung der Klassifikationsleistungen um 1-2%. Die besten Leistungen zeigten die Klassifikatoren SVM mit polynomialem Kern und MLP mit einer GER von 99,6%. Die optimale Anzahl der Haupt-

komponenten liegt im Bereich von 12 bis 14 Hauptkomponenten. Neben der Optimierung des Spektralteils des Hybriddatensatzes gibt es eine weitere Möglichkeit, die Leistungen zu verbessern, unter Anwendung von Merkmalsselektionsverfahren auf dem Bildteil in Kombination mit einer Hauptkomponentenanalyse auf dem Spektralteil. Dies lässt höhere Erkennungsraten für alle Klassen bzw. Materialien im Vergleich zur alleinigen Anwendung dieser Methode zu. Die damit erzielbaren Erkennungsraten unter Anwendung von svmPoly-Klassifikator auf dem Datensatz mit 5 Oberklassen liegen bei 100% für alle Klassen außer Klasse 2 (99%). Auf dem Datensatz mit 8 Materialklassen zeigt der Klassifikator ebenfalls sehr hohe Leistungen, nur einzelne, nicht mit 100%-iger Erkennungsrate detektierbare Materialien sind Beton (99,9%) und Leichtbeton (99,8%).

Die erreichten Ergebnisse zeigen, dass die Bild- und Spektralinformationen zusammen genutzt werden sollten, um eine sichere Erkennung von Bauschuttrezyklaten zu realisieren. Obwohl die alleinige Anwendung der Informationen bereits hohe Erkennungsraten (über 95%) erreichen lässt, ist eine sichere, dem Standard entsprechende Erkennung mit ER über 99% nur unter Anwendung von der fusionierten Information möglich. Die Hybrid-Information ermöglicht alle Klassen mit Ausnahme von Granit besser zu unterscheiden. Die Erkennung von schwer erkennbaren Klassen wie Leichtbeton und Beton wurde um 0,8-2% verbessert.

Insgesamt ist die entwickelte Methodik fähig, die Analyse der Bauschuttrezyklate mit Erkennungsraten von über 98,7% für Materialien und über 99% für Oberklassen durchzuführen. Hierbei werden die Anforderungen der DIN Norm 4226-100 für die Erkennung von Baustoffkomponenten erfüllt.

16. Ausblick

Die durchgeführten Untersuchungen haben gezeigt, dass die Lösung einer komplexen Aufgabe wie die Bauschutterkennung ein großes Forschungsfeld darstellt. Weitere Untersuchungen der entwickelten Methode sind notwendig. Es existieren noch weitere Ansätze, welche im Rahmen dieser Dissertation noch nicht ausführlich untersucht bzw. beschrieben werden konnten. Die Untersuchungen dieser Arbeit wurden auf eine begrenzte Anzahl an Bauschuttmaterialien durchgeführt. Der Lerndatensatz sollte zukünftig noch um neuentwickelte Baumaterialien (Kompositmaterialien) ergänzt werden. Die Variantenvielfalt der bereits verwendeten Proben sollte ebenfalls noch weiter erhöht werden sowie der Umfang des Datensatzes vergrößert werden.

Auf der Hardware-Seite stellt die Anwendung von ebenfalls ortsaufgelösten Hyperspektralkameras eine weitere Optimierungsmöglichkeit dar, welche im Rahmen des Projektes RezykDetect bereits im Ansatz untersucht wurde. Ortsaufgelöste Hyperspektralkameras können nicht nur die ortsaufgelöste Bildinformation, sondern auch die ortsaufgelöste Spektralinformation in mehreren Kanälen aufnehmen und für eine Klassifikation zur Verfügung stellen. Das erlaubt potenziell neben der Erkennung einfacher Materialien zusätzlich auch die Erkennung komplexer Mischungen unterschiedlicher Materialien in einem Baustoff (z.B. Hybrid-Gesteinskörnungen mit mehreren chemischen Materialanteilen, wo z. B.: ein Material mit einem anderen Material ummantelt vorliegt).

Des Weiteren sollte die Anwendbarkeit neuester Algorithmen aus dem Bereich des maschinellen Lernens, wie z. B.: das Deep Learning, ergänzend für einige Komponenten der entwickelten Erkennungsroutine untersucht werden. So könnten z. B.: faltende neuronale Net-

ze (convolutional neural network (CNN)) für die Merkmalsextraktion direkt ohne zusätzliche Software (wie z. B. Halcon) verwendet werden. Die Ausgangsschicht des Netzes kann mit einem Klassifikator verbunden werden. Die Vergrößerung der Probenanzahl ermöglicht die Anwendung von Deep-Learning-Verfahren, was zu einer weiteren Verbesserung der Leistungen führen könnte.

Abbildungsverzeichnis

Tabellenverzeichnis

Verzeichnis häufig verwendeter Formelzeichen und Abkürzungen

Formlezeichen - Griechische Buchstaben

α	Lernschritt
α_i	Lagrange-Multiplikator
β	Winkel im Bildraum HSI
η	Lernrate neuronales Netzes
θ_i	Koeffizient logistischer Regression
λ	Wellenlänge
$\hat{\mu}$	Mittelwert
ξ_i	Schlupfvariable
$\xi(\cdot)$	Netzeingabefunktion
ρ	Korrelationskoeffizient
σ_{RBF}	Konstante für Gauß-Kern
σ_x	Quadratwurzel der Varianz der Variable x
$\hat{\sigma}$	mittlere Varianz
$\phi(x)$	Merkmalstransformationfunktion
χ^2	Chi-Quadrat-Statistik
ψ	Bewertung der Leistungen
ω_j	Kategorie

Formlezeichen - Lateinische Buchstaben

a	Konstante für Sigmoid- und Polynomkern
A	Objektfläche
A_{kur}	Fläche der kleinsten umschreibenden Rechtecks
Acc_i	einzelne Genauigkeit
Acc	Genauigkeit
b	Schwellwert
B	Verzweigungsgrad
$Blue$	Blau
c_i	Klassenlabel
C_i	Cost Parameter
Cov	Kovarianz
$\mathbf{D}$	Datensatz
$\mathbf{D}_i$	Teildatensatz
E_i	Anzahl der Beobachtungen
$\nabla E(\cdot)$	Fehlerfunktion
fp	Anzahl falsch positiver Klassifikationen
fn	Anzahl falsch negativer Klassifikationen
g_i	Unähnlichkeitsmaß
$g(x, y)$	multivariate Verteilung
$Green$	Grün
$gt(x)$	Ground-Truth-Label

$h(\cdot)$	Aktivierungsfunktion
$H(\cdot)$	Entropie
Hue	Farbton
$i(\cdot)$	Entropie-Unreinheitsmaß
$I(\cdot)$	Indikatorfunktion
I_i	Intensitätswert des Spektrums
$I_{i(zentriert)}$	zentrierte Intensitätswert des Spektrums
$\bar{I}$	Mittelwert des Spektrums
$IG(\cdot)$	Informationsgewinn
$Intensity$	Intensität
$J(\mathrm{w})$	Rayleigh-Koeffizient
J_M	Relevanzkriterium
$\mathbf{K}$	Anzahl der Feldern
$k(x)$	vorhergesagtes Label
$K(x_i, x_j)$	Kernfunktion
$l(\theta)$	logarithmische Likelihood-Funktion
$\mathcal{L}(\theta)$	Maximum-Likelihood-Funktion
L_{dual}	doppelte Lagrange-Zielfunktion
L_p	Lagrange-Strafe
m_i	Mittelwert
$\tilde{m}_i$	Projektion des Mittelwertes
N	Bezeichnung des Knotenpunktes
$nearHit$	nearest Hit
$nearMiss$	nearest Miss
net	Aktivierung
O_i	Anzahl der Beobachtungen
$p(\cdot)$	Verteilung
$P(\cdot)$	Wahrscheinlichkeit
q	Wahrscheinlichkeit
r	interner Parameter SVM
R	Rechteckförmigkeit
R^N	Merkmalsraum
r_{ky}	Korrelationskoeffizient
r_{kk}	Kreuzkorrelationskoeffizient
Red	Rot
rn	Anzahl richtig negativer Klassifikationen
rp	Anzahl richtig positiver Klassifikationen
$\tilde{s}_i$	Varianz der projizierten Daten
S_B	Interklassendistanzmatrix
S_i	Varianzmatrix
S_k	Satz der k-nächsten Nachbarn
S_W	Intraklassenvarianmatrix
$Saturation$	Sättigung
T	Transponierte
t_i	Temperatur
u	Instanz

v	Distanzparameter
$V(j)$	Anzahl der Stimmen
w	Gewichtsvektor
X	Menge aller möglichen Merkmale
x_i	spezifischer Merkmalswert
x_k	Koordinate des Objektes
x_s	Schwerpunktskoordinate des Objektes
Y	Menge aller möglichen Merkmale
$\widehat{y}(z)$	vorhergesagtes Klassenlabel
y_i	spezifischer Merkmalswert
y_k	Koordinate des Objektes
y_s	Schwerpunktskoordinate des Objektes
z_i	Ausgabe eines Neurons

Allgemeine Abkürzungen

Abb.	Abbildung
ca.	cirka
d.h.	das heißt
etc.	et cetera
u.a.	und andere
usw.	und so weiter
z.B.	zum Beispiel
z.T.	zum Teil

Fachspezifische Abkürzungen

AVIRIS	NASA Airborne Visible/Infra-Red Imaging Spectrometer
CART	Classification and Regression Trees
CCD	Charge Coupled Device
CFS	Correlation-based Feature Selection
CMOS	Complementary Metal Oxide Semiconductor
CNN	Convolutional Neural Network
DIN	Deutsche Industrie Norm
DNN	Deep Neural Network
EER	Einzelerkennungsrate
ELM	Extreme Learning Machine
FTIR	Fourier Transform Infra-Red
GER	Gesamterkennungsrate
HSI	Hue, Saturation, Intensity (Bezeichnung eines Farbraums)
ID3	Iterative Dichotomiser 3
IR	Infrared (Infrarot)
JHU	Johns Hopkins University
JPL	Jet Propulsion Laboratory
kNN	k-Nearest Neighbors
KNN	künstliche neuronale Netz
LD	lineare Diskriminanzkomponente

LDA	Lineare Diskriminanzanalyse
LED	Light-emitting diode
LogitBoost	Logistische Regression mit Boosting
MLP	mehrlagiges Perzeptron
NB	Naive Bayes
NIR	Near-infrared (Nah-Infrarot)
NN	Nearest-Neighbor
PC	Principal component
PCA	Principal Component Analysis (Hauptkomponentenanalyse)
RBF	radiale Basisfunktion
RF	Random Forest
RFE	Recursive Feature Elimination
RGB	Red, green, blue (Bezeichnung eines Farbraums)
ROC	Receiver-Operating-Characteristic
SBE	Sequential backward elimination
SFS	Sequential forward selection
SLP	Single Layer Perceptron (SLP)
SVM	Support-Vektor-Maschine
svmLinear	Support-Vektor-Maschine mit linear Kern
svmPoly	Support-Vektor-Maschine mit polynomial Kern
svmRadialSigma	Support-Vektor-Maschine mit Gaussian (RBF) Kernel
USGS	United States Geological Survey
UV	Ultraviolett
VIS	Visible spectrum (sichtbares Licht)

Literaturverzeichnis

[Abmayr, 1994] Abmayr, W. (1994). *Einführung in die digitale Bildverarbeitung*. Vieweg+Teubner Verlag.

[Adams, 2004] Adams, M. (2004). *Chemometrics in analytical spectroscopy*. RSC analytical spectroscopy monographs. Royal Society of Chemistry, 2nd ed edition.

[Aggarwal, 2014] Aggarwal, C. C. (2014). *Data Classification: Algorithms and Applications*. Chapman and Hall/CRC Data Mining and Knowledge Discovery Series. Chapman and Hall/CRC, 1 edition.

[Aggarwal, 2015] Aggarwal, C. C. (2015). *Data Mining: The Textbook*. Springer International Publishing, 1 edition.

[Anding, 2010] Anding, K. (2010). *Automatisierte Qualitätssicherung von Getreide mit überwachten Lernverfahren in der Bildverarbeitung*. Dissertation, Technische Universität Ilmenau.

[Anding et al., 2013] Anding, K., Garten, D., and Linß, E. (2013). Application of Intelligent Image Processing in the Construction Material Industry. In: Acta Imeko Journal 2013, Vol. 2, Number 1, p. 61 – 73.

[Anding et al., 2011] Anding, K., Linß, E., Träger, H., Rückwardt, M., and Göpfert, A. (2011). Optical Identification of Construction and Demolition Waste by Using Image Processing and Machine Learning Methods. In *Joint International IMEKO TC1+ TC7+ TC13 Symposium*.

[Baldridge et al., 2009] Baldridge, A., Hook, S., Grove, C., and G. Rivera, G. (2009). The ASTER spectral library version 2.0. *Remote Sensing of Environment*, 113:711–715.

[Baudelet, 2014] Baudelet, M. (2014). *Laser spectroscopy for sensing: Fundamentals, techniques and applications*. Woodhead Publishing Series in Electronic and Optical Materials 43. Woodhead Publishing.

[Bergstra and Bengio, 2012] Bergstra, J. and Bengio, Y. (2012). Random search for Hyper-parameter Optimization. *Journal of Machine Learning Research*, 13:281–305.

[Bernath, 2005] Bernath, P. F. (2005). *Spectra of atoms and molecules*. Topics in Physical Chemistry). Oxford University Press, 2nd ed edition.

[Bishop, 2006] Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Information science and statistics. Springer, 1st ed. 2006. corr. 2nd printing edition.

[Breiman, 1996] Breiman, L. (1996). Bagging Predictors. *Machine Learning*, 24(2):123–140.

[Breiman, 2001] Breiman, L. (2001). Random Forests. *Machine Learning*, 45(1):5–32.

[Burger and Burge, 2015] Burger, W. and Burge, M. (2015). *Digitale Bildverarbeitung: Eine algorithmische Einführung mit Java*. X.media.press. Springer Vieweg, 3 edition.

[Chapelle et al., 2002] Chapelle, O., Vapnik, V., Bousquet, O., and Mukherjee, S. (2002). Choosing Multiple Parameters for Support Vector Machines. *Machine Learning*, 46(1-3):131–159.

[Ciresan et al., 2012] Ciresan, D. C., Meier, U., Masci, J., and Schmidhuber, J. (2012). Multi-column deep neural network for traffic sign classification. *Neural Networks*, 32:333–338.

[Cortes and Vapnik, 1995] Cortes, C. and Vapnik, V. (1995). Support-Vector Networks. *Machine Learning*, 20(3):273–297.

[Craven, 2010] Craven, P. (2010). Washed recycled aggregates. online picture, Flickr, https://www.flickr.com/photos/cdeimages/4865299925/. Zugriff am: 07.10.2018.

[DAfStb, 2010] DAfStb (2010). Deutscher Ausschuss für Stahlbeton: DAfStb-Richtlinie Beton nach DIN EN 206-1 und DIN 1045-2 mit rezyklierten Gesteinskörnungen nach DIN EN 12620.

[Daintith, 2008] Daintith, J. (2008). *Dictionary of Chemistry, 6th Edition*. Oxford University Press, USA, 6 edition.

[Dal-Canton, 2011] Dal-Canton, S. (2011). Konzeption eines mechanischen Systems zur bildanalytischen Beurteilung verschiedener Gesteinsarten der Korngröße 8 bis 16 mm. Master's thesis, Technische Universität Ilmenau, Ilmenau.

[Demant et al., 2011] Demant, C., Streicher-Abel, B., and Springhoff, A. (2011). *Industrielle Bildverarbeitung: Wie optische Qualitatskontrolle wirklich funktioniert, 3. Auflage*. Springer, 3., aktualisierte aufl. edition.

[DIN 1045-2, 2008] DIN 1045-2 (2008). Tragwerke aus Beton, Stahlbeton und Spannbeton - Teil 2: Beton - Festlegung, Eigenschaften, Herstellung und Konformität - Anwendungsregeln zu DIN EN 206-1.

[DIN 206-1, 2001] DIN 206-1 (2001). Beton - Teil 1: Festlegung, Eigenschaften, Herstellung und Konformität; Deutsche Fassung EN 206-1:2000.

[DIN 4226-100, 2002] DIN 4226-100 (2002). Gesteinskörnungen für Beton und Mörtel – Teil 100: Rezyklierte Gesteinskörnungen.

[DIN 4226-101, 2017] DIN 4226-101 (2017). Rezyklierte Gesteinskörnungen für Beton nach DIN EN 12620 - Teil 101: Typen und geregelte gefährliche Substanzen.

[DIN 4226-102, 2017] DIN 4226-102 (2017). Rezyklierte Gesteinskörnungen für Beton nach DIN EN 12620 - Teil 102: Typprüfung und Werkseigene Produktionskontrolle.

[DIN EN 12620, 2008] DIN EN 12620 (2008). Gesteinskörnungen für Beton; Deutsche Fassung EN 12620:2008.

[Do et al., 2007] Do, C. B., Foo, C.-S., and Ng, A. Y. (2007). Efficient Multiple Hyperparameter Learning for Log-linear Models. In *Proceedings of the 20th International Conference on Neural Information Processing Systems*, NIPS'07, pages 377–384, USA. Curran Associates Inc.

[Dougherty, 2013] Dougherty, G. (2013). *Pattern Recognition and Classification*. Springer.

[Duda et al., 2001] Duda, R. O., Hart, P. E., and Stork, D. G. (2001). *Pattern classification*. Wiley, 2nd ed edition.

[Friedman et al., 2009] Friedman, J., Hastie, T., and Tibshirani, R. (2009). *The Elements Of Statistical Learning, Data Mining Inference And Prediction*. Springer, 2 edition.

[Garten et al., 2013] Garten, D., Anding, K., Lerm, S., Linss, G., and Brückner, P. (2013). Image Analysis of Natural Products. OCM 2013, Karlsruhe, S. 157-167.

[Gasteiger and Engel, 2003] Gasteiger, J. and Engel, T. e. (2003). *Chemoinformatics*. John Wiley & Sons.

[Gonzalez and Woods, 2007] Gonzalez, R. C. and Woods, R. E. (2007). *Digital Image Processing*. 3rd Edition. Prentice Hall, 3 edition.

[Goodfellow et al., 2016] Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep Learning*. MIT Press. `http://www.deeplearningbook.org`.

[Grossberg, 1973] Grossberg, S. (1973). Contour enhancement, short term memory, and constancies in reverberating neural networks. *Studies in Applied Mathematics*, 52(3):213–257.

[Guyon and Elisseeff, 2006] Guyon, I. and Elisseeff, A. (2006). *Feature Extraction: Foundations and Applications*. Studies in Fuzziness and Soft Computing 207. Springer-Verlag Berlin Heidelberg, 1 edition.

[Hall, 1998] Hall, M. A. (1998). *Correlation-based Feature Subset Selection for Machine Learning*. PhD thesis, University of Waikato, Hamilton, New Zealand.

[Heaton, 2015] Heaton, J. (2015). *Artificial Intelligence for Humans, Volume 3: Neural Networks and Deep Learning*. CreateSpace Independent Publishing Platform.

[Hinton et al., 2012] Hinton, G., Deng, L., Yu, D., Dahl, G. E., Mohamed, A. r., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T. N., and Kingsbury, B. (2012). Deep Neural Networks for Acoustic Modeling in Speech Recognition: The Shared Views of Four Research Groups. *IEEE Signal Processing Magazine*, 29(6):82–97.

[Hollas, 2004] Hollas, J. M. (2004). *Modern spectroscopy*. J. Wiley, 4th ed edition.

[Hollstein et al., 2017] Hollstein, F., Wohllebe, M., Herling, M., Cacho, I., and Arnaiz, S. (2017). Sorting of construction and demolition waste by hyperspectral-imaging. In *International HISER Conference on Advances in Recycling and Management of Construction and Demolition Waste*, pages 33–36.

[Hsu et al., 2003] Hsu, C.-W., Chang, C.-C., and Lin, C.-J. (2003). A practical guide to support vector classification. Technical report, Department of Computer Science, National Taiwan University.

[Huang, 2015] Huang, G.-B. (2015). What are extreme learning machines? Filling the gap between Frank Rosenblatt's dream and John von Neumann's puzzle. *Cognitive Computation*, 7(3):263–278.

[Huang and Chen, 2007] Huang, G.-B. and Chen, L. (2007). Convex incremental extreme learning machine. *Neurocomputing*, 70(16):3056–3062.

[Huang et al., 2004] Huang, G.-b., Zhu, Q.-y., and Siew, C.-k. (2004). Extreme learning machine: a new learning scheme of feedforward neural networks. In *2004 IEEE International Joint Conference on Neural Networks (IEEE Cat. No.04CH37541)*, volume 2, pages 985–990 vol.2.

[Hulley et al., 2017] Hulley, G., Hook, S., Fisher, J., and Lee, C. (2017). ECOSTRESS, a NASA Earth-Ventures Instrument for studying links between the water cycle and plant health over the diurnal cycle. In *2017 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, pages 5494–5496.

[Irvine and Pollack, 1968] Irvine, W. M. and Pollack, J. B. (1968). Infrared optical properties of water and ice spheres. *Icarus*, 8(1):324 – 360.

[Jähne, 2012] Jähne, B. (2012). *Digitale Bildverarbeitung*. Springer Berlin Heidelberg.

[Jolliffe, 2002] Jolliffe, I. (2002). *Principal Component Analysis*. Springer, 2nd edition.

[Kasun et al., 2013] Kasun, L., Zhou, H., Huang, G.-B., and Vong, C.-M. (2013). Representational learning with elms for big data. *IEEE Intelligent Systems*, 28:31–34.

[Kelemen et al., 2008] Kelemen, A., Abraham, A., and Liang, Y. (2008). *Computational Intelligence in Medical Informatics*. Springer Publishing Company, Incorporated, 1st edition.

[Kreislaufwirtschaft Bau, 2017] Kreislaufwirtschaft Bau (2017). Mineralische Bauabfälle Monitoring 2014. Bundesverband Baustoffe – Steine und Erden e. V., Berlin.

[Krizhevsky et al., 2012] Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). ImageNet Classification with Deep Convolutional Neural Networks. In *Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'12, pages 1097–1105, USA. Curran Associates Inc.

[Kruse et al., 2015] Kruse, R., Borgelt, C., Braune, C., Klawonn, F., Moewes, C., and Steinbrecher, M. (2015). *Computational Intelligence: Eine methodische Einfuehrung in Kuenstliche Neuronale Netze, Evolutionaere Algorithmen, Fuzzy-Systeme und Bayes-Netze*. Computational Intelligence. Springer Vieweg, 2 edition.

[Kuhn, 2009] Kuhn, M. (2009). The caret Package.

[Laws, 1980] Laws, K. I. (1980). Rapid Texture Identification. In *Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series*, volume 238 of , pages 376–380.

[Linß, 2014] Linß, E. (2014). Entwicklung eines Recyclingverfahrens für Mauerwerksbaustoffe auf Basis hyperspektraler Nahinfrarot-Sensorik. Zwischenbericht zum AiF-Forschungsprojekt KF3033025, Bauhaus-Universität Weimar.

[Linss et al., 2012a] Linss, E., Anding, K., Schnellert, T., and Ludwig, H.-M. (2012a). Identification of Construction and Demolition Waste by Using Image Processing in the Visual and Near-infrared Spectrum, and Machine Learning Methods. Internationale Baustofftagung: 18. ibausil, Weimar, Tagungsbericht, Band 2, S. 1026-1033.

[Linss et al., 2017] Linss, E., Karrasch, A., and Landmann, M. (2017). Sorting of mineral construction and demolition wastes by near-infrared technology. In *International HISER Conference on Advances in Recycling and Management of Construction and Demolition Waste*, pages 29–32.

[Linss et al., 2010] Linss, E., Müller, A., Escher, M., and Anding, K. (2010). Qualitätsparameter von Recyclingbaustoffen. Fachtagung Recycling R10.

[Linss et al., 2012b] Linss, E., Traeger, H., Ludwig, H.-M., Anding, K., and Linss, G. (2012b). Study of the Identification of Aggregates of Construction and Demolition Waste by Using Object Recognition Methods. IALCCE 2012: 3rd International Symposium on Life-Cycle Civil Engineering, Vienna, Austria.

[Liu and Motoda, 2008] Liu, H. and Motoda, H. (2008). *Computational methods of feature selection*. Chapman and Hall/CRC data mining and knowledge discovery series. Chapman and Hall/CRC.

[McCallum and Nigam, 1998] McCallum, A. and Nigam, K. (1998). A comparison of event models for Naive Bayes text classification. In *Workshop on Learning for Text Categorization*, pages 41–48. AAAI Press.

[Mehta and Monteiro, 2005] Mehta, P. and Monteiro, P. (2005). *Concrete: Microstructure, Properties, and Materials*. McGraw-Hill Professional, 3 edition.

[Minsky and Papert, 1969] Minsky, M. and Papert, S. (1969). *Perceptrons*. MIT Press, Cambridge, MA.

[Mohri et al., 2012] Mohri, M., Rostamizadeh, A., and Talwalkar, A. (2012). *Foundations of Machine Learning*. Adaptive Computation and Machine Learning. The MIT Press.

[Murphy, 2012] Murphy, K. P. (2012). *Machine Learning: A Probabilistic Perspective*. The MIT Press.

[Nawy, 2008] Nawy, E. (2008). *Concrete construction engineering handbook*. CRC Press, 2nd ed edition.

[Nizar et al., 2008] Nizar, A. H., Dong, Z. Y., and Wang, Y. (2008). Power Utility Nontechnical Loss Analysis with Extreme Learning Machine Method. *IEEE Transactions on Power Systems*, 23(3):946–955.

[Robnik-Šikonja and Kononenko, 2003] Robnik-Šikonja, M. and Kononenko, I. (2003). Theoretical and Empirical Analysis of ReliefF and RReliefF. *Mach. Learn.*, 53(1-2):23–69.

[Runkler, 2012] Runkler, T. (2012). *Data Analytics: Models and Algorithms for Intelligent Data Analysis*. Vieweg+Teubner Verlag, 1 edition.

[Runkler, 2010] Runkler, T. A. (2010). *Data Mining: Methoden und Algorithmen intelligenter Datenanalyse*. Vieweg+Teubner Verlag, 1 edition.

[Sabins, 2007] Sabins, F. (2007). *Remote Sensing: Principles and Applications, Third Edition*. Waveland Press.

[Savitzky and Golay, 1964] Savitzky, A. and Golay, M. J. E. (1964). Smoothing and Differentiation of Data by Simplified Least Squares Procedures. *Analytical Chemistry*, 36:1627–1639.

[Süße and Rodner, 2014] Süße, H. and Rodner, E. (2014). *Bildverarbeitung und Objekterkennung: Computer Vision in Industrie und Medizin*. Vieweg+Teubner Verlag, 1 edition.

[Sermanet et al., 2013] Sermanet, P., Kavukcuoglu, K., Chintala, S., and Lecun, Y. (2013). Pedestrian Detection with Unsupervised Multi-stage Feature Learning. In *Proceedings of the 2013 IEEE Conference on Computer Vision and Pattern Recognition*, CVPR '13, pages 3626–3633, Washington, DC, USA. IEEE Computer Society.

[Settle, 1997] Settle, F. A. (1997). *Handbook of Instrumental Techniques for Analytical Chemistry*. Prentice Hall PTR.

[Siesler et al., 2002] Siesler, H. W., Ozaki, Y., Kawata, S., and Heise, H. M. (2002). *Near-Infrared Spectroscopy: Principles, Instruments, Applications*. Wiley-VCH, 1 edition.

[Skoog et al., 2006] Skoog, D. A., Holler, F. J., and Crouch, S. R. (2006). *Principles of Instrumental Analysis sixth edition*. Brooks Cole, 6 edition.

[Skoog et al., 2013] Skoog, D. A., West, D. M., Holler, F. J., and Crouch, S. R. (2013). *Fundamentals of Analytical Chemistry*. Cengage Learning, 9 edition.

[Sun et al., 2008] Sun, Z.-L., Choi, T.-M., Au, K.-F., and Yu, Y. (2008). Sales forecasting using extreme learning machine with applications in fashion retailing. *Decision Support Systems*, 46(1):411 – 419.

[Witten and Frank, 2011] Witten, I. H. and Frank, E. (2011). *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann, San Francisco, 3rd edition.

[Witten et al., 2016] Witten, I. H., Frank, E., Hall, M. A., and Pal, C. J. (2016). *Data Mining, Fourth Edition: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 4th edition.

[Xanthopoulos et al., 2013] Xanthopoulos, P., Pardalos, P., and Trafalis, T. (2013). *Robust data mining*. Springer briefs in optimization. Springer.

[Zaki and Meira Jr., 2014] Zaki, M. J. and Meira Jr., W. (2014). *Data Mining and Analysis: Fundamental Concepts and Algorithms*. Cambridge University Press, draft edition.

[Zhu and Hastie, 2004] Zhu, J. and Hastie, T. (2004). Classification of gene microarrays by penalized logistic regression. *Biostatistics*, 5(14):427–443.

Anhang

Liste der verwendeten Merkmale für die Bilddatensätze:

- area_center
- circularity
- FormfaktorKompaktheit
- ContLength
- Convexity
- DiameterMaxAbstand2erRandpunkte
- Anisometry
- Bulkiness
- StructureFactor
- Ra
- Rb
- Ellipsen_Verhaeltnis
- RadiusgroessterInnenkreis
- M11normiert
- M20normiert
- M02normiert
- PHI1relativnormiert
- PHI2relativnormiert
- M21
- M12
- M03
- M30
- PSI1invariantesgeometrischesZentralmoment
- PSI2invariantesgeometrischesZentralmoment
- PSI3invariantesgeometrischesZentralmoment
- PSI4invariantesgeometrischesZentralmoment
- mittlererAbstand
- SigmamittlereAbweichungvomAbstand
- RoundnessVerhaeltnisMittelwertzuStandard-abweichung
- AnzahlderPolygonstuecke
- Sehnenanzahl
- KFactorAbweichungderSehnenanzahlvomQuadrat
- LFactormittlereSehnenanzahlproZeile
- mittlereLaengederSehnen
- Bytes
- RadiusdeskleinstenumschliessendenKreises
- Length1
- Length2
- Verhaeltnis
- AreaXLD
- ContrastH
- LengthXLD
- RadiusXLDKreis
- Radius1XLDEllipse
- Radius2XLDEllipse
- EllipsenVerhaeltnisXLD
- MinDist
- MaxDist
- AvgDist
- SigmaDist
- Length2XLD
- cM33
- cM32
- cM31
- cM30
- cM23
- cM22
- cM21
- cM20
- cM13
- cM12
- cM11
- cM03
- cM02
- c1
- c2
- c3
- c4
- c5
- c6
- c7
- AreaS
- AreaH
- AreaI
- EnergyS
- CorrelationS
- HomogenityS
- ContrastS
- EnergyH
- CorrelationH
- HomogenityH
- Beta1I

EnergyI
CorrelationI
HomogenityI
ContrastI
RaS
RbS
PhiS
RaH
RbH
PhiH
RaI
RbI
PhiI
EntropyS
AnisotropyS
EntropyH
AnisotropyH
EntropyI
AnisotropyI
AlphaS
BetaS
GammaS
AlphaH
BetaH
GammaH
AlphaI
BetaI
GammaI
Alpha1S
Beta1S
Gamma1S
DeltaS
EpsilonS
ZetaS
Alpha1H
Beta1H
Gamma1H
DeltaH
EpsilonH
ZetaH
Alpha1I
Deviation1H
Deviation1I
MeanH_ll
MeanI_ll
DeviationH_ll
Gamma1I
DeltaI
EpsilonI
ZetaI
FuzzyEntropyS
FuzzyEntropyH
FuzzyEntropyI
FuzzyPerimeterS
FuzzyPerimeterH
FuzzyPerimeterI
MeanS
DeviationS
MeanH
DeviationH
MeanI
DeviationI
MinS
MaxS
RangeS
MinH
MaxH
RangeH
MinI
MaxI
RangeI
MRowS
MColS
Alpha2S
Beta2S
Mean1S
MRowH
MColH
Alpha2H
Beta2H
Mean1H
MRowI
MColI
Alpha2I
Beta2I
Mean1I
Deviation1S
MeanI_rs
DeviationH_rs
DeviationI_rs
MeanH_rr
MeanI_rr

DeviationI_ll
MeanH_lr
MeanI_lr
DeviationH_lr
DeviationI_lr
MeanH_ee
MeanI_ee
DeviationH_ee
DeviationI_ee
MeanH_es
MeanI_es
DeviationH_es
DeviationI_es
MeanH_ew
MeanI_ew
DeviationH_ew
DeviationI_ew
MeanH_se
MeanI_se
DeviationH_se
DeviationI_se
MeanH_ss
MeanI_ss
DeviationH_ss
DeviationI_ss
MeanH_sw
MeanI_sw
DeviationH_sw
DeviationI_sw
DeviationI_re
DeviationH_rr
DeviationI_rr
MeanH_rw
MeanI_rw
DeviationH_rw
DeviationI_rw
MeanH_wl
MeanI_wl
DeviationH_wl
DeviationI_wl
MeanH_we
MeanI_we
DeviationH_we
DeviationI_we
MeanH_ws
MeanI_ws
DeviationH_ws
DeviationI_ws
MeanH_wr
MeanI_wr
DeviationH_wr
DeviationI_wr
DeviationH_re
MeanH_rl
MeanI_rl
DeviationH_rl
DeviationI_rl
MeanH_re
MeanI_re
MeanH_rs

Thesen

1. Die Abbruchabfälle werden größtenteils im Erd- und Straßenbau wegen der hohen Heterogenität des aufbereiteten Bauschutts verwendet (Downcycling). Die Qualität der Gesteinskörnungen als Betonzuschlagsstoff hat einen großen Einfluss auf die späteren Betoneigenschaften.

2. Die Realisierung einer automatisierten Bauschutterkennung ist eine hochkomplexe Aufgabe, welche eine Vielzahl an entsprechenden Anpassungen der Algorithmen des maschinellen Lernens, der Spektroskopie und der Bildverarbeitung benötigt, um eine qualitativ gute, den Normen im Bereich der Betonherstellung genügende Lösung erzielen zu können.

3. Die Bauschuttproben zeigen eine geringe Interklassenvariabilität bei einer gleichzeitig relativ großen Intraklassenvariabilität. Deswegen spielt eine sinnvolle Datensatzstrukturierung für das Erreichen guter Erkennungsleistungen eine große Rolle.

4. Die Zusammenfassung in der Bauschutt-Subklassen in größere Gruppen (Oberklassen) soll vor einem stattfindenden Klassifikatortraining durchgeführt werden, da dies zu einer Erhöhung der Generalisierung und zu einer Verringerung des Rechenaufwandes im Vergleich zu einer späteren Zusammenfassung nach dem Training führt.

5. Textur- und Farbmerkmale weisen einen Haupteinfluss auf die Klassifikationsleistung für den Bilddatensatz auf.

6. Die Anwendung von Filter-Merkmalsselektionsverfahren auf den Bildmerkmalen verbessert die Gesamterkennungsraten um $0{,}1-0{,}5\%$ für den Bilddatensatz. Hier ist die Anwendung des InfoGain-Merkmalsselektionsverfahrens und des chiSquare-Verfahrens optimal für die Klassifikationsaufgabe geeignet, aufgrund des geringen Rechenaufwandes bei gleichzeitig bester erreichbarer Leistung.

7. Nur einige Bauschuttmaterialien zeigen charakteristische spektrale Eigenschaften im VIS-Bereich, die für die Lösung der Klassifikationsaufgabe angewendet werden können. Im NIR-Bereich ($1100-2100$ nm) weisen die Bauschuttmaterialien mehr Eigenschaften auf, die für eine Trennung relevant sind.

8. Die Merkmalstransformation mittels der Hauptkomponentenanalyse (PCA) verbessert die Leistungen auf dem Spektraldatensatz drastisch. Im Gegensatz dazu verschlechtern sich die Leistungen auf dem Bilddatensatz bei Einführung komplexerer Merkmale mittels Hauptkomponentenanalyse.

9. Eine direkte Anwendung der Klassifikatoren auf dem Hybriddatensatz ohne Merkmalsselektion führt zu schlechteren Ergebnissen. Der Grund dafür ist der negative Einfluss spektraler Merkmale aufgrund ihrer sehr starken Korrelation untereinander.

10. Die Anwendung der am besten geeigneten Merkmalsselektions-/Merkmalsextraktionsverfahren (InfoGain-Filter für den Bilddatensatz und Hauptkomponentenanalyse für den Spektraldatensatz) in Kombination auf entsprechenden Teilen des Hybriddatensatzes lässt bestmögliche Leistungen erreichen.

11. Die Bild- und Spektralinformationen sollten zusammen genutzt werden, um eine sichere, dem Standard entsprechende Erkennung von Bauschuttrezyklaten realisieren zu können. Eine alleinige Verwendung von Bildmerkmalen oder Spektralmerkmalen führt nicht zu gleich hohen Erkennungsleistungen wie die gemeinsame Anwendung der Hybridinformation (bildanalytisch und spektral).

Eigene wissenschaftliche Veröffentlichungen

1. Latyev, S., Voronin, A., Anding, K., Linß, E., Kuritcyn, P.: *Optical-Electronic Methods and Means for Identification of Substances and Materials*, In: Scientific and Technical Journal Priborostroenie – 2013, Vol. 56, Number 10., p. 81-87

2. Kuritcyn, P., Anding, K., Linß, E., Latyev, S. M.: *Increasing the Safety in Recycling of Construction and Demolition Waste by Using Supervised Machine Learning*, In: Journal of Physics: Conference Series 588 (2015) 012035, 2015

3. Anding, K., Kuritcyn, P., Garten, D.: *Using artificial intelligence strategies for process-related automated inspection in the production environment*, In: Journal of Physics: Conference Series 772 012026, Volume 772, Number 1, 2016

4. Anding, K., Kuritcyn, P., Linß, E., Latyev, S. M.: *Significant Characteristics in VIS- and IR-Spectrum of Construction and Demolition Waste for High-precision Supervised Classification*, 2nd International Conference on Optical Characterization of Materials, OCM 2015, Karlsruhe, 2015

5. Trambitckii, K.; Kuritcyn, P., Garten, D., Anding, K., Polte, G.: *Metal surface quality assessment using 2D texture features*, 14th IMEKO TC10 Conference on Technical Diagnostics New Perspectives in Measurements, Tools and Techniques for system's reliability, maintainability and safety Milan, Italy, 27.-28. Juni 2016

6. Anding, K., Kuritcyn, P., Garten, D.: *Using artificial intelligence strategies for process-related automated inspection in the production environment*, IMEKO TC1-TC7-TC13 Joint Symposium 2016 University of California, Berkeley 3.-5. August 2016

7. Linß, E., Garten, D., Karrasch, A., Kuritcyn, P., Andig, K.: *Analyseverfahren zur automatisierten Qualitätssicherung für rezyklierte Gesteinskörnungen auf Basis hyperspektraler Bildinformationen im VIS und NIR*, In: Tagungsband der 20. Internationalen Baustofftagung ibausil, F.A. Finger-Institut für Baustoffkunde, Bauhaus-Universität Weimar, 2018, S. 2-373 - 2-380

8. Kuritcyn, P., Anding, K., Linß, E., Notni, G.: *Using hybrid information of colour image analysis and SWIR-spectrum for high-precision analysis of construction and demolition waste*, 4th International Conference on Optical Characterization of Materials, OCM 2019, Karlsruhe, 2019 [eingereicht, angenommen]